# Using Statistical and Machine Learning Methods to Improve Treatment Success in Patients with Schizophrenia

Agbedjro, Deborah

*Awarding institution:*
King's College London

# Using Statistical and Machine Learning Methods to Improve Treatment Success in Patients with Schizophrenia

**Deborah Agbedjro**

**A thesis submitted for the degree of Doctor of Philosophy in Biostatistics**

**King's College London**
**Institute of Psychiatry, Psychology & Neuroscience**
**Department of Biostatistics**

# Contents

# List of Tables

# List of Figures

# Abstract of thesis

**Background:** People with schizophrenia (SCZ) suffer from impaired cognitive abilities and these are associated with poor functional outcomes. Cognitive Remediation Therapy (CRT) has been shown effective in improving the cognitive deficits of SCZ. Because there is evidence for CRT treatment heterogeneity of outcomes, there is a need to identify CRT predictors of differential response using moderation analysis of high dimensional psychiatric data, which typically contain relatively large percentages of missingness. This will contribute to precision medicine treatment, understanding mechanism responsible of differential therapy responses, and better prognosis.

**Aims:** The primary aim of this PhD consisted of developing a CRT precision medicine model, using computer intensive statistical learning methods able to deal with high dimensional psychiatric data containing large percentages of missingness in the predictors and smaller percentages in the outcome. Secondary aims were overcoming the following problems: variable selection or measurement of variable importance in the model, multicollinearity and overfitting, and summarising commensurate outcomes in one latent outcome.

**Methods:** A simulation study comparing four statistical learning methods (Lasso, Elastic-net, Random Forests and Conditional Inference Random Forests) combined with two missing data imputation techniques (Multivariate Imputation using Chained Equations and MissForest) was run. The combined methods were assessed according to their optimism-corrected (via bootstrap internal validation) prediction accuracy and variable selection performance in different scenarios. The best method was chosen to develop a CRT precision medicine model using individual participant data from seven randomised controlled trials with approximately 400 patients. Factor scores from a latent summary measure of cognitive commensurate outcomes, obtained via Factor Analysis, was used as the model dependent variable, to accommodate the above univariate statistical learning methods.

**Results**: In the simulations, the method combining MissForest imputation with Lasso was

the best compromise between prediction accuracy and clinical interpretability. MissForest-Lasso was then used to develop an internally validated precision medicine model, which selected only a weak moderator of treatment response. The model was therefore mainly prognostic.

**Conclusion:** In future research, more modalities of data, such as genetics, OMICS and neuroimaging data, are recommended to successfully identify moderators of CRT success.

# Acknowledgements

I am really thankful to my first supervisor Daniel Stahl for his constant paternal help, kindness, availability, professional advice and guidance during this three year and a half PhD.

Additionally, I would like to thank my second supervisors Artemis Koukounari and Matteo Cella for their present support and statistical and psychological expertise respectively. Further gratitude goes to the Department of Biostatistics in King's College, in particular to Cédric Ginestet, Silia Vitoratou, Nicholas Magill, Andrew Pickles and many others for their competent assistance and friendship.

Moreover, I would like to thank Daniel Stamate from the Department of Computing in Goldsmith College University of London for providing me with computing facilities suitable to my analyses during the project.

Finally, I would like to thank my family and my friends for their lovely care and psychological help. A huge thanks goes to God, Who helped me through all the mentioned people and allowed me to successfully complete this project.

# List of abbreviations

| | |
|---|---|
| AQT | Ammons Quick Test |
| BPRS | Brief Psychiatric Rating Scale |
| DoCTRS/DCTRS | Database of Cognitive Training and Remediation Studies |
| CATFLU | Category fluency |
| CART | Classification and Regression Trees |
| CCD | CIRcuiTS Combined Data |
| CFA | Confirmatory Factor Analysis |
| CRT | Cognitive Remediation Therapy |
| CVLT | California Verbal Learning Test |
| DSM | Diagnostic and Statistical Manual of Mental Disorders |
| EFA | Exploratory Factor Analysis |
| ES | Effect size |
| FAS | Verbal fluency test using the letters F, A, and S |
| FP(s) | Fake predictor(s) or noise variables |
| FPR | False positive rate of selection |
| FS | Factor scores |
| GLM | Generalised linear model |
| GLMM | Generalised linear mixed models |
| HAY | Hayling test |
| HCQOL | Heinrichs-Carpenter Quality of Life Scale |
| ICD | International Classification of Diseases |
| i.i.d. | independent and identically distributed |
| IQ | Intelligence quotient |
| LFA | Longitudinal Factor Analysis |
| LNS | Letter-Number Span |
| MAR | Missing at random |
| MARS | Multivariate Adaptive Regression Splines |
| MCAR | Missing completely at random |
| MNAR | Missing not at random |
| MICE | Multivariate imputation using chained equations |
| MSET | Modified six elements task |
| NRMSE | normalized root mean squared error |

| | |
|---|---|
| PANSS | Positive and Negative Syndrome Scale |
| PFC | Proportion of falsely classified entries |
| RCT | Randomised Controlled Trial |
| RMSEA | Root Mean Square Error of Approximation |
| PPV | Positive predictive value of selection |
| RSE | Rosenberg Self-esteem Scale |
| TMTA | Trailmaking test part A |
| TMTB | Trailmaking test part B |
| TP(s) | True predictor(s) |
| S1 | Simulation scenario 1: no missing data, no assumption of moderation |
| S2 | Simulation scenario 2: missing data (MCAR and MAR), no assumption of moderation, complete outcome |
| S3 | Simulation scenario 3: no missing data, assumption of moderation |
| S4 | Simulation scenario 4: missing data (MCAR and MAR) in predictors, assumption of moderation, complete outcome |
| S5 | Simulation scenario 5: missing data in predictors and outcome (MCAR and MAR), assumption of moderation |
| S6 | Simulation scenario 6: missing data in predictors, interactions in the imputation model, complete outcome (only for MICE-Lasso) |
| SANS | Scale for the Assessment of Negative Symptoms |
| SAPS | Scale for the Assessment of Positive Symptoms |
| SAS-II | Social Adjustment Scale II |
| SBS | Social Behaviour Scale |
| SCZ | Schizophrenia |
| SEN | Sensitivity of selection |
| TAU | Treatment as usual |
| RF | Random Forests |
| WAIS | Wechsler Adult Intelligence Scale |
| WCST | Wisconsin Card Sorting Test |

# Chapter 1

# Introduction

This thesis will pursue statistical learning methods and missing data imputation techniques in order to develop a precision medicine model, predicting treatment outcome heterogeneity among patients with schizophrenia (SCZ) treated with Cognitive Remediation Therapy (CRT).

This chapter will introduce the concept of SCZ and the intervention CRT as one of its psychological treatments. Precision medicine with the identification of moderators of treatment will be explained. Then, I will discuss how statistical learning methods can overcome some limitations of classical statistical methods in the analysis of large psychiatric datasets. In a second part, I will introduce the methods to develop prediction models and their performance assessment through discrimination, calibration and validation. As missing data constitute an important problem in mental health studies, imputation techniques will be presented alongside approaches combining them with statistical learning methods. Finally, dimension reduction techniques, such as factor analysis, will be considered in order to deal with multiple outcomes that measure the same construct. The aim and specific objectives of the thesis will ultimately be defined.

## 1.1 Literature review

### 1.1.1 Schizophrenia and cognitive remediation therapy (CRT)

Schizophrenia is a severe and debilitating mental health condition. Worldwide, one in every 100 people will develop SCZ, but only about 10% of these will achieve complete symptom remission (Jaaskelainen et al. 2013). Despite a significant investment in both pharmacological and psychosocial interventions, the majority of the people affected will have long term disability and not meet personal and professional life goals. Schizophrenia is generally considered as a

disorder with poor prognosis. A study investigating 18 prospective long-term studies found that less than 50% of patients had poor outcomes and, equally, less than 50% had good outcomes (Van Os and Kapur 2009).

Schizophrenia is a disorder characterized mainly by positive, negative and cognitive symptoms (Van Os and Kapur 2009):

- **Positive symptoms** or *psychosis* are irrational thoughts and feelings 'added on' to a person that reflect an excess or distortion of normal function. The following are positive symptoms:
  - *Delusions*: abnormal beliefs and convictions (persecutory delusions are known to be the most common),
  - *Hallucinations*: abnormal sensory experiences (auditory, visual and tactiles),
  - *Disorganized speech*: derailment, lack of association, use of newly coined words while speaking,
  - *Catatonic behaviour*: muscular rigidity and tightness or hyperactivity;
- **Negative symptoms** refer to a decrease or absence of normal function, resulting in reduced motivation:
  - *Avolition*: lack of motivation, drive,
  - *Anhedonia*: inability to experience pleasure,
  - *Asociality*: inability and unwillingness to socialise,
  - *Reduced emotional intensity and reactivity*;
- **Cognitive symptoms** are a diminution in neurocognitive abilities:
  - *Attention deficit*,
  - *Slower information processing: lacking insight and understanding, problem solving difficulties*,
  - *Memory deficit* (working and long term memory),
  - *Low executive functioning*: difficulty with organization and following directions.

Historically the treatment emphasis has been on tackling positive symptoms. Nevertheless, there is increasing recognition that cognitive deficits are associated with the illness' functional problems (Cella, Huddy, et al. 2012), such as the ability to live in the community, work, function in a social environment and the quality of life (Rajji, Miranda, and Mulsant 2014). It is

suggested that treating cognitive deficits in patients improves general outcomes of SCZ and prognosis. Cognitive abilities or cognition are terms used to identify thinking skills such as attention, information processing speed, working and long term memory, executive functioning and ability to plan and regulate behaviour. Cognitive abilities in people with SCZ are typically one standard deviation below the general population (Fioravanti et al. 2005). Cognitive deficits negatively affect recovery and vocational functioning (in particular social cognitive problems, M. Green and Harvey 2014).

The two main important intervention modalities for people with SCZ are pharmacological and psychosocial. Pharmacological treatments largely rely on the use of anti-psychotic medications. There is evidence that this treatment has an effect on positive symptoms, but has little or no effect on cognitive and negative symptoms (Van Os and Kapur 2009). There is evidence that psychosocial therapies also successfully reduce positive symptoms; for example, cognitive behavioural therapy for psychosis reduces distress and negative affect related to psychosis but cognitive deficits are not addressed (Morrison 2001). In contrast, cognitive remediation therapy (CRT) is a psychological intervention targeting specifically the cognitive symptoms of SCZ. CRT is defined as "a behavioural training based intervention that aims to improve cognitive processes (e.g. attention, memory, executive function, processing speed, social cognition and metacognition) with the goal of durability and generalization" (Wykes, Huddy, et al. 2011). CRT uses *drill and practice*, i.e. repetitive exercises and intensive training to improve cognition over time. CRT has been shown in more than 40 randomised controlled trials to be beneficial and cost-effective in reducing the burden of cognitive problems in people with SCZ, even after controlling for sources of bias like unmasked assessment (Wykes, Huddy, et al. 2011). This treatment is delivered in combination with pharmacological or psychological treatments for positive symptoms. There are two approaches for CRT (Cella, Huddy, et al. 2012):

- *drill and practice*: consists of gradually more challenging exercises without a specific procedure to follow, trial and error or implicit learning strategies.
- *drill and practice plus strategy*: identifies the particular cognitive deficit and applies an explicit learning strategy in everyday life.

Although there is evidence for the effectiveness of CRT, the current CRT approach does not tailor the most suitable intervention for an individual. In fact, there was evidence of heterogeneity in the effect sizes of CRT across trials for the 40 study meta-analysis by Wykes, Huddy, et al. 2011. A way to improve the efficacy of this intervention is trying to explain the heterogeneity of treatment outcome through *precision medicine* (see next Subsection 1.1.2).

### 1.1.2 Prediction modelling for precision medicine

Clinical trials support the efficacy of CRT for people with SCZ in improving cognition, however there is variability in outcomes. Wykes, Huddy et al. (2011) in their 40-study meta-analysis reported that cognitive effect sizes ranged from -0.24 to 2.35 (29 studies with non significant effect size) with overall effect size of 0.45 (95% confidence interval: 0.31-0.59). However, the authors did not find any variables explaining the heterogeneity of cognitive effects. Therefore, identifying patients who can benefit from treatment and key predictor variables of treatment success is essential for improving treatment efficacy. This concept is increasingly referred to as *precision* or *personalised* or *stratified medicine* or care, which aims to tailor intervention to particular groups of patients (PROGRESS, Hingorani et al. 2013). Precision medicine is also used for diagnosis and risk assessments (Redekop and Mladsi 2013). 'Precision medicine' and 'personalised medicine' are terms used interchangeably, but 'precision medicine' is more appropriate as the aim is to identify the best approach of the intervention for a particular person, while 'personalised' could be misunderstood to think that treatments are developed for each person (*Help me understanding genetics: Precision medicine* 2018). It is fundamental to develop prediction models from patients' characteristics such as demographics, clinical, genetic and psychological variables (i.e. baseline variables, measured before treatment), which can predict likely treatment outcome. Such models provide guidance for clinical decision making and help clinicians to recommend the best treatment approach. Those patients' characteristics or biomarkers, that predict which treatment will be optimally suited for a patient, are called **moderators**.

In statistical terms, a moderator is a variable that affects the strength of the relationship between a dependent and independent variable. In this case, it amplifies or weakens the relationship between treatment type and outcome (VanderWeele 2015), it identifies *whom* or *under what conditions* treatment works (Kraemer, Wilson, et al. 2002). A moderator predicts differential treatment response, but it does not need to be a prognostic factor for the outcome. However, there are moderators that are also predictors. A moderator, by definition, precedes treatment (or what it moderates), that in turn precedes the outcome. This means that a moderator needs to be measured before treatment, i.e. it is a baseline variable independent from the treatment variable. Moderation can be also defined as effect modification by baseline variables (Dunn et al. 2015). An effect modifier is a variable for which the effect of an intervention on the outcome differs across the variable levels, like a moderator, but it is not necessarily a baseline variable (VanderWeele 2009). An effect modifier can also be a post randomization variable, for example a process variable (see next Subsection 1.1.3).

Prediction modelling is used to identify moderators of treatment heterogeneity which reliably predict future outcome of new unseen patients (Steyerberg 2009). Identifying moderators of treatment effect is typically done using regression-based approaches that assess the effect on the outcome of statistical interaction terms (Dunn et al. 2015) given by the product of baseline variables and treatment type (i.e. new treatment and treatment as usual or no treatment, Kraemer, Frank, and Kupfer 2006). However, it is important here to clarify that the significant effect of a statistical interaction term in a regression model might not identify effect modification and could instead be evidence of biological interaction or both. Biological interaction between the effects of two exposures (interventions) on the outcome in causal inference (counterfactual framework) happens when the effect of the exposures on the outcome is different from the combination of the two effects considered separately (VanderWeele 2009). The presence of statistical interaction depends on the scale of the measurement used, for example there may be significant statistical interaction on the risk difference scale (the most common) and not on the risk ratio scale or the odds ratio scale (Kupper and Hogan 1978). In this thesis, I will refer to statistical interaction to identify effect modification by baseline variables with the plain term 'interaction'.

The gold standard study, that can identify and distinguish moderators of a treatment, is the randomised controlled trial (RCT). It is the most effective way to find moderators in study data. The key characteristics of an RCT is the random allocation of patients in treatment and control groups (Sibbald and Roland 1998). An RCT consists of having a control or comparison treatment in the study, which avoids misleading effects like regression to the mean or drift in measurement or confounding. An RCT is a planned experiment to evaluate the benefits of one or more treatments, usually for patients with a specific medical condition. The main reason to do RCTs is to produce comparable treatment and control groups by equally distributing confounders between the two groups, and therefore to obtain unbiased estimates of treatment effect. A well-designed trial provides the most rigorous method for evaluating efficacy (or effectiveness) of treatments and safety (BS Everitt and Wessely 2009). In particular, RCTs need to be:

- *Controlled*: any intervention needs to be compared to one or more other interventions; in drug therapy trials, there will often be a placebo control or another active condition which is regularly used in clinical practice; in psychological treatments, the control condition will often be treatment as usual (TAU).

- *Unbiased*: there needs to be a fair comparison between the treatments, with no bias

whether deliberate or accidental; thus, randomisation is crucial, i.e. patients are randomly allocated to a particular treatment group such that the allocation cannot be predicted in advance and the obtained results are unbiased.

- *Large and appropriately powered*: in order to obtain a precise estimate and to balance measured and unmeasured confounders of any treatment effect, sufficiently large numbers are required. Also, having a large sample size lowers the risk of overfitting the data, when analysed with a statistical model.

Typically, evidence-based medicine with continuous outcome uses RCTs to evaluate a marginal effect of the treatment of interest (i.e. the average effect of treatment on the population). On the other hand, precision medicine focuses on the differences between individuals and aims to predict differential treatment response (Leon 2012).

Traditional evidence-based medicine tries to infer causal relationships between covariates and outcome. In recent years, there has been a need to shift from 'explaining' towards 'predicting' treatment outcome for new individuals, and precision medicine better suits this new goal (Shmueli 2010). Predictive modelling tries to minimize model bias and sampling variance at the same time through validation techniques (see Subsection 1.2.1), in order to accurately predict response in new data. Instead, explanatory modelling aims to minimize model bias and then sampling variance to theoretically predict (infer) an association between covariate and outcome using the same dataset used to develop the model. In prediction modelling the theoretical model itself is typically not of interest and, sometimes, an accurate prediction model can be seen as a 'black box', lacking clinical interpretability (see next Subsection 1.1.4 about statistical learning). Consequently, the best explanatory model may differ from the best predictive model, and predictive power cannot be drawn from explanatory power (Stahl and Pickles 2018). On the contrary, prediction modelling is a promising tool also for explanatory research to generate new theories, to act as explorative data analysis and to allow the comparison of competing theories and the identification of new patterns (Shmueli 2010).

### 1.1.3 Moderation of schizophrenia CRT treatment

Despite the effectiveness of CRT for the cognitive symptoms of SCZ, there is considerable heterogeneity in treatment success among patients. However, little is known about why some patients respond better to treatment than others. Thus, identifying potential moderators of CRT may help to tailor general treatment to individuals better.

So far, moderators for CRT have been investigated in single studies and in traditional (aggregated) meta-analyses of CRT effects on cognitive performance and on functioning. McGurk et al. (2007) was a 26-study meta-analysis that showed evidence for the following methodological quality moderators on functioning: studies that provided adjunctive psychiatric rehabilitation had stronger effect sizes than studies without psychiatric rehabilitation; similarly, CRT programs that used drill plus strategy were more effective than drill and practice only CRT. Also age was found to moderate CRT with younger patients reacting better to treatment than older patients (McGurk et al. 2007). Wykes, Huddy et al. (2011), a 40-study meta-analysis, could replicate all the moderation findings of McGurk et al. (2007) apart from identifying age as moderator on functioning. However, moderators of CRT on cognition have not been found yet.

Adjunctive psychiatric rehabilitation and type of therapy (drill and practice vs drill plus strategy) were considered as moderators of treatment in the literature. However, they cannot be considered as such in a causal inference framework as they only refer to the active treatment and are not applicable to the control treatment. These variables that measure an aspect of a specific treatment are called process variables (Dunn et al. 2015). Therapy alliance, compliance, treatment strategy are all examples of such variables. By definition a process variable is an intermediate outcome of therapy and so a conditional variable, i.e. can only be observed for the people who receive the treatment of interest. The mechanism investigated by McGurk et al. (2007) and Wykes, Huddy et al. (2011) would be called 'effect modification by a post-randomisation variable'. The term 'treatment effect moderation' is usually restricted to effect modification by baseline (pre-treatment) variables (see previous Subsection 1.1.2).

Suggestions that study age (years since publication date), therapy duration, computer presentation, presence of adjunctive psychiatric rehabilitation, and type of therapy moderate CRT on cognitive outcome have been put forward by Wykes, Huddy et al. (2011) in their meta-analysis study using meta-regression. However, these still need to be confirmed or re-tested because their meta-analysis suffered from missingness by design (the variables measured in each study were not the same for all studies). Putative patient characteristic moderators that were proposed were: age (McGurk et al. 2007, Wykes, Reeder, Landau, Matthiasson, et al. 2009) and symptoms (Garety et al. 2008). Also, Cella, Huddy, et al. 2012) reviewed the literature about moderators for CRT on cognition in single RCTs and advanced a number of potential moderators to be further investigated including age, severity of baseline cognitive deficits and level of psychopathology.

Although there have been attempts to understand heterogeneity of CRT success through successful moderation analyses (McGurk et al. 2007), some of the results could not be repli-

cated for lack of data (Wykes, Huddy, et al. 2011) and, to my knowledge, no precision medicine model for CRT was developed so far. In particular, moderators of CRT on cognition have not been found yet. Therefore there is still need to identify and confirm moderators of CRT by analysing **multiple RCTs' data**. Using individual participant data meta-analysis, instead of meta-regression of aggregated effect sizes and sample sizes only, may reduce reporting biases, improve the generalisability of the results to populations and increase statistical power (Riley et al. 2013, Curran and Hussong 2009, Higgins et al. 2001).

The data available for this PhD project are individual participant data from seven RCTs for CRT (see Subsection 3.1.1). Because of the small number of studies for which we have accessible data, an individual patient meta-regression analysis cannot be done. This is one of the reasons why statistical learning methods will be considered (see next Subsection 1.1.4).

### 1.1.4 High dimensional data and statistical learning

Assessing moderators of CRT on cognitive outcome for people with SCZ involves a large number of variables measured in relatively small samples: measures of baseline cognitive abilities, medications, demographics, symptoms and other variables need to be included as predictors in the statistical analysis together with all their interactions with the treatment variable. All these potential predictor values form a **high dimensional dataset**, i.e. a dataset characterized by few dozen to many thousands of continuous variables and/or levels of categorical variables not necessarily measured on a relatively large sample size (*Clustering High-Dimensional Data* 2012).

Classical statistical models (e.g. linear regression models) are not suitable to cope with a number of potential predictors close to or higher than the number of observations. Model selection methods (such as stepwise selection) are often based on inclusion or exclusion of a variable depending on its p-value or similar procedures, such as Akaike's Information Criterion (AIC) or Bayesian Information Criterion (BIC). These procedures result in overestimating treatment effects (Tibshirani 1996) and in **overfitting**: the larger the number of variables relative to sample size is, the higher the variance of the coefficient estimates will be (Harrell 2001, James et al. 2013). Subsequently, results can often not be replicated (Ioannidis 2005, Nuzzo 2014, Stahl, Pickles, et al. 2012). Also, in high dimensional datasets, classical methods cannot address the problem of **multicollinearity**. This arises when there are two or more predictor variables in a model that are highly correlated, resulting in high variance in prediction.

In recent years, computer intensive **Statistical Learning Methods** (SLMs), including machine learning, are increasingly used to overcome the described problems of classical infer-

ential methods (Hastie, Tibshirani, and Friedman 2008). Statistical learning unifies classical statistics (which makes probabilistic assumptions about underlying phenomena) and machine learning (highly computer intensive algorithms for prediction). SLMs typically focus on prediction accuracy and less on inference (see previous Subsection 1.1.2).

Classical methods develop and test their models on the same sample, but the error in predicting new cases obtained this way is underestimated and models are unlikely to generalize well. If a given method applied on the same training sample returns a small prediction error (*training error*), but a large error when applied on new data (*test error*), the model overfits the data. A key aspect of statistical learning is the use of internal validation to assess model fit on unseen (hold-out) data, using cross-validation and bootstrapping procedures to overcome overfitting (see Subsection 1.2.1).

We can distinguish between *supervised learning*, i.e. the model building process is guided by the presence of the outcome, and *unsupervised learning*, i.e. we only observe the input data without any measurements of the outcome and we try to recognise the data patterns or clusters in the data (James et al. 2013). This project only focuses on supervised learning methods.

In supervised statistical learning, we assume that there is a relationship between a quantitative outcome $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ and a matrix of features (predictors) as column vectors $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$, according to an unknown function $f$: $\boldsymbol{y} = f(\boldsymbol{X}) + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is a random error term, which is independent of $\boldsymbol{X}$, such that $E(\boldsymbol{\epsilon}) = \underline{\mathbf{0}}$. Supervised statistical learning models are developed according to statistical decision theory. For this theory, the error in prediction of $\boldsymbol{y}$ using an approximation of $f$, $\hat{f}$, measured by the *loss functions* $L(\boldsymbol{y}, \hat{f}(\boldsymbol{X}))$, is minimised by penalizing it in order to have a criterion for choosing the most suitable predictive function. We can then rewrite all statistical learning methods as minimization problems of "Loss + Penalty". The penalty function depends on *tuning parameters* which are non-negative and need to be determined through resampling methods like cross-validation and the bootstrap (see Subsection 1.2.1), in order to minimise the out-of-sample error: this process is called *tuning*.

Key concepts in supervised statistical learning are the trade-off between model prediction accuracy and model interpretability and the bias-variance trade-off (Hastie, Tibshirani, and Friedman 2008, James et al. 2013):

- **Prediction accuracy-interpretability trade-off**: in general, if the interpretability of a method is low, then its prediction accuracy and hence flexibility are high. For example, linear regression is a relatively restrictive and inflexible method, because its function represents only a simple hyperplane (or line in the case of simple linear regression with one covariate). However, it is highly clinically interpretable. Other methods (such as support

vector machine, which we will see below) are more flexible because their functions generate complicated shapes in space and can accommodate the data well with high prediction accuracy, but at the same time they are difficult to interpret. Depending on the analysis one needs to run, the suitable trade-off between prediction accuracy and interpretability is chosen (James et al. 2013).

- **Bias-variance trade-off**: the variance of a statistical learning method $\hat{f}$ is the variability of $\hat{f}$ if we estimated it using a different data set. The bias of a statistical learning method refers to the error between the true $f$, which may be extremely complicated, and $\hat{f}$, a simpler approximation of $f$. As the flexibility of a method increases, the bias tends to initially decrease faster than the variance increases. However, at a certain point, increasing flexibility does not influence much the bias, which levels off, but starts to significantly increase the variance (Hastie, Tibshirani, and Friedman 2008). The expected prediction error (or extra-sample error) of a model can be expressed as the following sum: "irreducible error (i.e. $Var(\boldsymbol{\epsilon})$)+ bias$^2$+variance". The aim is to minimise the expected prediction error by finding the best compromise between bias and variance. As the sum of squared bias and variance is the mean squared error (MSE), it is sufficient to minimise the MSE. This is done by assessing the model at hold-out data with different tuning parameters and selecting the tuning parameter giving the model with the minimum MSE.

A brief overview of supervised statistical learning methods (James et al. 2013) follows:

- **Regularised regression methods** (RMs): such methods aim to avoid overfitting and thus improve prediction accuracy of regression models when the number of predictors (features) is large relative to sample size. RMs reduce variance of estimation at the cost of some bias which is induced by introducing a constraint on the magnitude of the regression coefficient parameter estimates. The coefficients for RMs are estimated by minimising the residual sum of squares (loss function) plus a penalty monitored by a set of tuning parameters, which controls the strength of the regularization. As a consequence, the model coefficients are shrunk towards zero. Some RMs may shrink all coefficients relatively by the same amount without performing variable selection (e.g. Ridge L2 regularization), while other RMs like the 'Least Absolute Shrinkage and regression Operator' (LASSO, Tibshirani 1996, see Subsection 2.1.1) and Elasticnet (Zou and Hastie 2005, see Subsection 2.1.1) shrink some of the coefficients to be exactly zero resulting in **variable selection** and interpretability of the model. Also RMs with less used types of penalty are used for variable selection: the Dantzig Selector (performance similar to Lasso, Candes and

Tao 2007, Bickel, Ritov, and Tsybakov 2009), the Group-Lasso penalty (selects or omits groups of variables when potential predictors are structured into groups known a-priori, Yuan and Lin 2011), Adaptive Lasso (coefficients of strong predictors are shrunk less than coefficients of weak predictors in large samples, Zou 2006) and Smoothly Clipped Absolute Deviation (SCAD, similar performance to Adaptive Lasso, Fan and Li 2001). Most RMs are based on the Generalized Linear Models (GLM), while penalized Generalized Linear Mixed Models (GLMM) which allow the analyses of repeated measurements or other clustered data, are at early stages of development (Tutz and Groll 2011, Schelldorfer, Meier, and Bühlmann 2014, R package `lmmen`, 2017).

- **Non-linear models**: apart from the classical *polynomial regression* that extends simple linear models by adding powers of the predictor variable as new predictors, and their regularised versions, *step functions* (piece-wise constant functions) are used as non-linear models for qualitative variables. Another non-linear class of methods used in statistics are non-parametric *regression splines* methods (James et al. 2013). These are even more flexible than polynomials since the range of the matrix of explanatory variables is divided into different regions, in each of which a polynomial function is fit to the data and the obtained polynomial lines are joint smoothly at the region boundaries or *knots*. The regularised versions of regression splines are the *smoothing splines*: we penalize the residual sum of squares criterion and minimise it in order to get more smoothness and to avoid overfitting. In case we need to analyse multiple predictors with a non-linear model, *generalised additive models* (GAMs, Hastie, Tibshirani, and Friedman 2008) or *multivariate adaptive regression splines* (MARS, JH Friedman 1991) are used.

- $K$**-Nearest-Neighbours** (KNN, N. Altman 1992): given a new patient feature value, the non-parametric method KNN first identifies the $K$ patients in the training data that have the closest feature values to the new observation. Then it predicts the observation's outcome value with the mode of the nearest neighbours outcome in case of discrete data or their mean in case of continuous data. The number of neighbours, $K$, is a tuning parameter usually chosen through resampling methods. The method prediction accuracy is good, however the model is not easy to interpret. KNN is also used for missing data imputation (see Subsection 1.2.2).

- **Trees** (Breiman, JH Friedman, et al. 1984): Classification and regression trees (CART) are obtained by recursively partitioning the predictor space (i.e. the set of possible values for the covariates) into disjoint regions, according to an error minimisation criterion, and

fitting an interpretable non-parametric prediction model within each partition. A given test observation is predicted with the mean or the mode of the training data in the region to which it belongs. The obtained whole tree usually overfits the data because it is usually too complex. Instead, a smaller tree with fewer splits can have a reduced variance and better interpretation at the cost of some bias. Therefore, the whole tree is then pruned back (i.e the terminal branches are cut) using cross-validation (see Subsection 1.2.1) to minimise the test error. Prediction accuracy after pruning is usually not very good (Breiman, JH Friedman, et al. 1984). CART can impute missing data (see Subsection 1.2.2).

- **Random Forests** (RF, Breiman 2001, see Subsection 2.1.1): RF consists of averaging trees built on bootstrap samples drawn from the training set, in order to have a good prediction accuracy. The trees in a RF are decorrelated because only a random sample of predictors are chosen each time a split in a tree is considered. The number of variables selected at each split is a tuning parameter for RF. RF improves the error rate of trees at the cost of less interpretability. However, it can assess **variable importance** (Hastie, Tibshirani, and Friedman 2008). Like CART, RF is used for missing data imputation (see Subsection 1.2.2).

- **Support vector machines** (SVM, Cortes and Vapnik 1995): in classification problems, SVM finds the best separating hyperplane that minimises the classification error and maximises the geometric margin of classification. The data are separated according to their classes. There are also regression SVM (Smola and Schölkopf 1998).

- **Neural Networks** (NN, Kriesel 2007): these are machine learning models used in both regression and classification settings consisting of sums of non-linearly transformed linear models. They can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. In recent years, an extension of the NN, Deep Learning, became increasingly popular for the ability to learn by example, like humans naturally do. They are particularly suited to extremely large data sets such as web data and mobile healths data (Goodfellow, Bengio, and Courville 2016).

Table 1.1 (adapted from Hastie, Tibshirani, and Friedman 2008) contrasts the methods handling of missing values, robustness to outliers, transformation sensitivity, computational costs, accuracy and interpretability, showing that the gold standards for prediction are RF, SVM and NN even though they lack accessibility of information. However, RF provides a rank of importance for the variables, which gives some interpretability to the method. On the other hand,

decision trees are fairly good for interpretability, but perform quite poorly in prediction. A compromise is the Lasso (and similarly Elasticnet), which is a learning method that provides fairly accurate results and good extraction of information.

Clinicians often want to interpret the model to identify key components of prediction and understand underlying processes in order to improve or develop new treatments. Also, a prediction tool in clinical practice is often preferable when it only has a small number of variables, even if such a parsimonious model implies a small loss in prediction accuracy. In fact, measuring many variables can be costly and also clinicians do not want to overwhelm the patient with too many tests and questionnaires.

A serious and common problem in psychiatric studies is the presence of missing data, which largely depends on patients' attendance to assessments. Data analyses, which drop observations containing missing values, loose a lot of information and can result in biased estimates (see Subsection 1.2.2).

All statistical (machine) learning algorithms have their advantages and disadvantages. One common advantage is that almost no assumptions about the data are needed. RMs like Lasso and Elasticnet are usually helpful when the aim is getting parsimonious and interpretable models from fully pre-specified models (i.e. full models), without introducing selection bias. More-

Table 1.1: Comparison in terms of performance of some different statistical learning methods ($K$-NN=$K$-nearest neighbour, MARS=multivariate adaptive regression splines, SVM=support vector machine, NN=neural networks). Key: A=good, B=fair and C=poor. Adapted from Hastie et al. (2008).

| **Characteristic** | **Lasso** | **$K$-NN, kernels** | **MARS** | **Trees** | **Random forests** | **SVM** | **NN** |
|---|---|---|---|---|---|---|---|
| Natural handling of data of "mixed" type | C | C | A | A | A | C | C |
| Handling of missing values | C | A | A | A | A | C | C |
| Robustness to outliers in input space | B | A | C | A | A | C | C |
| Insensitiveness to monotone transformation of inputs | C | C | C | A | A | C | C |
| Computational cost (large $N$) | A | C | A | A | B | C | C |
| Ability to deal with irrelevant inputs | A | C | A | A | A | C | C |
| Ability to extract linear combinations of features | A | B | C | C | C | A | A |
| Interpretability | A | C | A | B | B | C | C |
| Predictive power | B | A | B | C | A | A | A |

over, RMs deal well with large number of variables relative to sample size and high collinearity. However, although RMs' variable selection will include almost all the true predictors, many noise variables will be also chosen, if their inclusion improve the bias-variance trade-off to get minimal error (Fan and Lv 2009). On the other hand, RF will not perform variable selection and models will need all the input variables to predict the outcome. However, RF can return a rank of importance in prediction strength for the variables. RF can be preferable when dealing with genetic data, such as genome wide association studies, which are easy to measure, or when generating hypothesis on potential predictors is the goal. RF will predict with low variance, but the estimates will be more biased than RMs' estimates (Breiman 2001). If non-linear trends or interactions are present, RF will automatically model them. When data are highly correlated, RF's bias is minimised. RF can be used to impute missing data although it induces some bias (Cutler et al. 2009). Similarly, KNN can impute missing data and is accurate in model predictions, but it results to be a total 'black box'. Moreover, datasets with a large number of variables relative to sample size lead to the inability for KNN to find nearby neighbours for a given observation, and therefore to an increase in prediction error (James et al. 2013). On the contrary, MARS are simple to understand and interpret and perform automatic variable selection. However, MARS do not handle multicollinearity well and prefer datasets with a low number of variables relative to the sample size. Finally, NN and SVM are good for their high predictive power and their excellent handling of very large datasets. However, they are complex and difficult to understand. NN and Deep Learning will suit particularly well web-data analyses and other large scale data.

Prediction accuracy often varies very little between statistical learning methods if the number of variables is not extremely large, i.e. does not exceed about half the sample size (Khondoker et al. 2013, Stahl, Pickles, et al. 2012). Moreover, statistical learning methods may perform better in prediction in slightly different populations (Hand 2006).

There is increasing literature that applies these methods to healthcare and mental health research (Hahn, Nierenberg, and Whitfield-Gabrieli 2017), in particular the following articles analysed psychiatric data :

- Stahl et al. (2012) applied classification RMs and SVM to the reanalysis of infant event-related potential (ERP) data with small sample size in comparison to the number of feature variables. The authors internally validated the methods through cross-validation (see Subsection 1.2.1) and both methods were able to separate above chance groups of infants according to their risk of a later diagnosis of autism. A discussed disadvantage of the study was the inability of the selected methods to perform variable selection.

- Koutsouleris et al. (2016) developed two internally and externally validated machine learning models to predict end-of-treatment and follow-up outcomes in patients with first-episode psychosis. Their method consisted of combining KNN missing data imputation with SVM in a repeated nested cross-validation analysis. The authors also validated the methods through leave-site-out validation as they analysed data from 44 studies (König et al. 2007 and Steyerberg and Harrell 2016 and see Subsection 1.2.1). Statistical significance was determined using permutation testing. Nevertheless, selection bias was induced in the model development as a stepwise forward variable selection process using SVM was adopted. Moreover, leave-site-out and external validation analyses were done by using only a selected set of variables determined by the model itself without using the full developed model.

- Ramsay et al. (2018) aimed to identify baseline predictors of cognitive improvement after receiving CRT using the LASSO with cross-validation tuning. However, only 10 potential predictors measured on a single study were considered. Furthermore, the model was not validated (see Subsection 1.2.1). Finally, statistical significance was inferred by regressing the outcome on the variables selected by the LASSO with subsequent selection bias.

Prediction modelling in psychiatry has got many challenges (Bzdok and Meyer-Lindenberg 2018 and Iniesta, Stahl, and McGuffin 2016) and the aim of this PhD research is to overcome some of these, namely, developing a precision medicine model with good prediction accuracy, ideally good clinical interpretability and able to handle the large number of variables in the available clinical data with substantial amounts of missing data. Therefore, RMs such as Lasso and Elasticnet for automatic feature variable selection, and RF, for ease of handling missing data and its variable importance rank will be considered.

## 1.2 Introduction to methods

In this second part of the introduction, I will describe the general steps of developing, assessing and validating prediction models. I will then introduce the concept of missing data and some applications of statistical learning methods combined with missing data imputation techniques. Finally, a statistical way to reduce dimensionality of multiple outcomes will be presented.

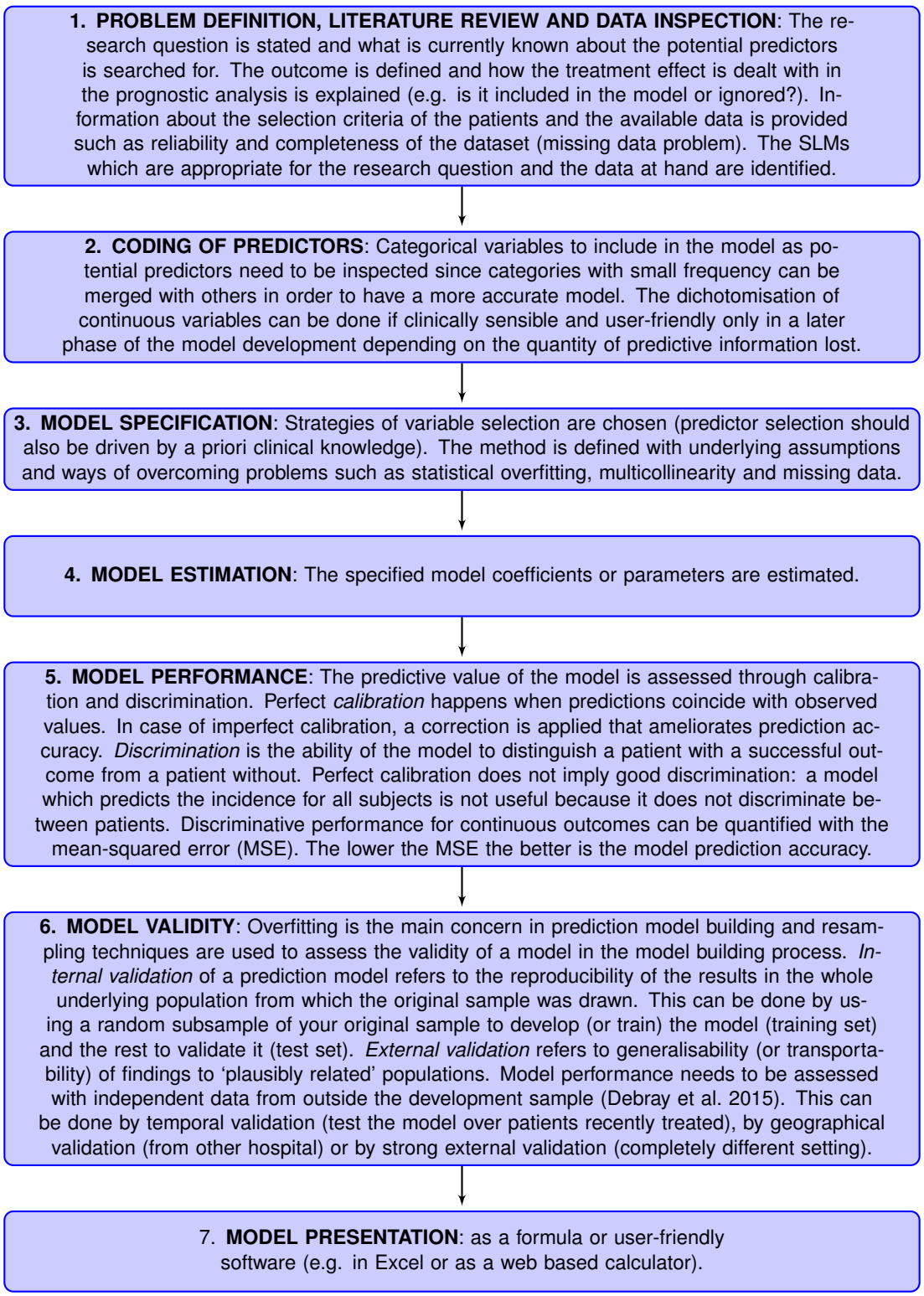
### 1.2.1 Prediction model performance measures and validation techniques

A good prediction model should perform well, i.e. be accurate in prediction and in separating groups of patients according to their risk, and be valid, i.e. able to make reliable predictions especially on unseen data (Steyerberg 2009).

Steyerberg and Vergouwe (2014) give clear guidelines on building prediction models, which consist of seven phases described in the flow chart 1.1. It is important that the prediction problem is well defined after having inspected the existing literature. The data to train and validate the model need to be checked in terms of reliability and completeness (missing data problem) and potential predictors need to be suitably coded. The model is then specified according to prior clinical knowledge, to avoid multicollinearity or overfitting depending on the method used. After estimating the model parameters, the model apparent performance and the model validated performance on new data (to evaluate replication of results) are assessed (steps 5 and 6 see Figure 1.1). Finally the developed validated model is then presented through a formula or a software function.

Performance measures are typically used to assess how well a model fits the same data, which it was trained with. However, these are not estimates of how well the model predicts new data points. It is then important to distinguish between performance on the training data (apparent performance), on hold-out data (internal performance) and on completely new data (external performance). In the next two paragraphs, I will first present prediction model performance measures and then explain how to assess the model performance correctly to obtain reliable measures of validity for the model.

Figure 1.1: Main steps for prediction modelling (Steyerberg and Vergouwe 2014)




### Performance

Step 5 in Figure 1.1 assesses the performance of the model. Measures of overall performance are the $R^2$, i.e. the amount of variability in outcomes that is explained by the prediction model

for continuous outcomes ($0 \leq \text{R}^2 \leq 1$) and the Brier Score (Brier 1950), i.e. the distance between observed and predicted probabilities, for binary outcomes. However, the $\text{R}^2$ estimator for the population is upwardly biased because it increases every time a predictor is added to the model without accounting for overfitting. Therefore it needs shrinkage. This is typically achieved in inferential statistics through the adjusted-$\text{R}^2$ version that considers the degrees of freedom used in the model.

The overall model performance quantifies how close predictions are to the actual outcome (using measures such as explained variation, $\text{R}^2$). Performance can further be evaluated in terms of discrimination and calibration:

- **discrimination**: the ability of the model to separate (discriminate) between high risk and low risk patients.
- **calibration**: the ability of the model to make unbiased estimates of the outcome, measuring the agreement between observed outcomes and predictions, e.g. if a prediction model assigned a 15% probability to develop a condition for each patient of a sample of 100 patients and 15 of them later developed the condition, the predictions would be calibrated (reliable).

Discrimination in classification is usually measured with the Area Under the Receiver Operating Characteristic (ROC) curve or C statistic or with the misclassification error. Instead, for continuous outcomes, the training **Mean Squared Error (MSE)** is used:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{y_i} - \hat{\boldsymbol{y}}_{\boldsymbol{i}})^2, \tag{1.1}$$

where $\boldsymbol{n}$ = sample size, $\boldsymbol{y_i}, i = 1, \ldots, n$ are the observed outcomes, $\hat{\boldsymbol{y}}_{\boldsymbol{i}}$ the predicted outcomes via the model $\hat{f}$.

The statistical learning version of the $\text{R}^2$ is the pseudo-$\text{R}^2 = 1 - \text{MSE}/Var(\boldsymbol{y})$. It differs from the classical regression $\text{R}^2$ as, for the pseudo-$\text{R}^2$, the variance of the outcome would be given by the total sum of squares divided by $n - 1$ and the MSE is given by the residual sum of squares divided by $n$. The classical $\text{R}^2$ has both denominators equal to $n$.

Calibration is assessed via the calibration slope $\beta$ and the calibration-in-the-large $\alpha$ obtained by

- regressing the $logit$ (i.e. the logarithm of the odds) of the observed probabilities that $\boldsymbol{y} = 1$

on the $logit$ of the probabilities that $\hat{\boldsymbol{y}} = 1$ in the case of *binary outcomes*:

$$logit(\boldsymbol{y} = 1) = \alpha + \beta logit(\hat{\boldsymbol{y}} = 1) \tag{1.2}$$

The above curve 1.2 is the one that better approximates the relationship between the predicted probabilities of the event and the true probabilities.

- regressing the observed outcome $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ on the predicted outcome $\hat{\boldsymbol{y}}$ in the case of *continuous outcomes*:

$$\boldsymbol{y} = \alpha + \beta\hat{\boldsymbol{y}} \tag{1.3}$$

The line returned by the above equation 1.3 is called the *calibration line* and it is the line that better approximates the relationship between the predictions and the observed values.

One way to quantify the unreliability of predictions is to measure what has to be done to make the calibration curve superimposed on the ideal curve, i.e. the 45 degree line.

When $\beta = 1$ and $\alpha = 0$, the model is perfectly calibrated and its calibration curve will be the 45 degree line. This happens when the model is an unbiased estimator of the outcome (e.g. GLM) and the calibration parameters are computed using the same training data used to build the model. The model is overfitting if $\beta < 1$, i.e. low predictions are too low and high predictions are too high, and underfitting if $\beta > 1$, when high predictions are too low and low predictions are too high (see Figure 1.2). The calibration-in-the-large $\alpha$ will be different from 0 to compensate.

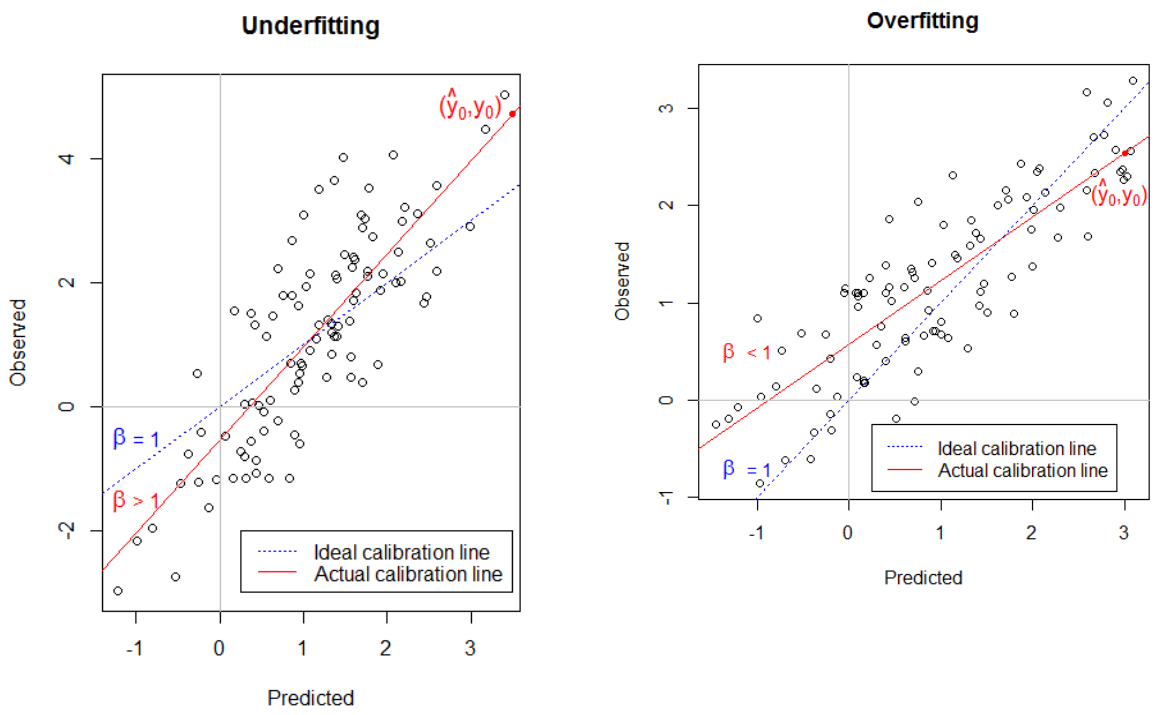
Good calibration does not imply good discrimination: the null model, without covariates, is well calibrated for definition, as the predicted risk is the same for all patients, but it is not a discriminative model distinguishing between severity of risks. Assuming the model has good calibration, a model which predicts risks with more variability has better discrimination.

### Validity

The sixth step of model building (see Figure 1.1) verifies the validity of the developed prediction model, assessed in the model building process through the model performance measures of calibration and discrimination, to ensure model generalizability to unseen data. Model validation addresses one of the major causes of model failure to predict well: overfitting. Overfitting leads to a too optimistic view of the developed model performance, since a model may predict well the data it was developed with (training data), but can fail in the prediction of new subjects from the underlying population. This optimism is the bias due to overfitting. **Optimism** is

Figure 1.2: Calibration lines for continuous outcomes in the cases of underfitting and overfitting. Observed outcomes are regressed on the model predicted outcomes. When the model underfits the data (Figure 1.2a), a high prediction $\hat{y}_0$ will be too low compared to the corresponding observed $y_0$: $\hat{y}_0 < y_0$. Vice-versa, when the model overfits the data (Figure 1.2b), a high prediction $\hat{y}_0$ will be too high compared to the corresponding observed $y_0$: $\hat{y}_0 > y_0$.

(a) Underfitting, $\beta > 1$ (b) Overfitting, $\beta < 1$




defined as true performance minus apparent performance, where true performance refers to the underlying population, and apparent performance refers to the estimated performance in the model development sample.

We can define three types of model validation:

- **Apparent validation**: testing the model on the whole original model development sample ($n$ data points);
- **Internal validation**: reproducibility of the model for the underlying population and setting where the development sample originated from;
- **External validation**: generalizability of the model to populations that are plausibly related, testing the model on new subjects.

Apparent validation is not a reliable method of assessing the predictive performance of a model. The performance measures of a prediction model are valid when they are evaluated on a test

sample which is independent from the training data. This can be done by internal and external validation. Internal validation is the estimate of replicability of the model using unseen samples from the same populations and it constitutes a minimum requirement to assess a model (Steyerberg 2009, Harrell 2001). External validation assesses the application of a model to different populations, e.g. the transportability to new settings, geographical and temporal, and is regarded as the gold standard to assess performance and clinical utility. In particular, when prediction models are developed on data from several studies, **internal-external validation** or **leave-site-out validation** (Steyerberg and Harrell 2016 and König et al. 2007) is recommended to estimate external validity of the model. Leave-site-out validation consists of leaving each study out in turn, developing the model on the remaining studies and testing it on the left out study. The final model is based on the pooled dataset and its validated performance is given by the average of the test performances.

Usually, independent data from new clinical populations for external validation are not available and the minimum requirement is to estimate the internal validity of a model. Internal validation can be done in several ways. The easiest procedure consists of developing the model on half or 2/3 of the sample (training set) and validating it on the rest of the data (test set) and this procedure is called **data-splitting or validation set**. However, to validate the model more accurately, a large sample needs to be available. **K-cross-validation** (CV) is repeated data-splitting and solves some of its problems: the data is divided into $k$ equal folds, then the first fold is omitted and the model is trained on the remaining $k-1$ folds to be then tested on the left-out fold and return an estimate of performance; the process is repeated $k$ times with the remaining folds and an average performance is returned that is nearly unbiased (biased upward, Borra and Di Ciaccio 2010). If $k = n$, then we will have a leave-one-out CV which minimises the bias, but has larger variance than $k$-CV (Hastie, Tibshirani, and Friedman 2008). The 5-10 CV is regarded as the best compromise between bias and variance (James et al. 2013 and Hastie, Tibshirani, and Friedman 2008). However, CV does not validate the model on the full size sample and thus the number of repetitions needed to achieve good estimates of performance often exceeds 200 (Harrell 2001). Therefore, if $k = 5$, the whole CV process will need to be repeated at least 40 times.

Another resampling method used for *model validation* is the **bootstrap** (Efron and RJ Tibshirani 1994, Harrell, Lee, and Mark 1996). In statistical inference the bootstrap is a robust statistical method used to quantify the uncertainty of an estimator (when we do not know its distribution), by mimicking the underlying population sampling process. In prediction modelling and statistical learning, bootstrap resampling is typically used to estimate prediction accuracy

of unseen data. Bootstrap resampling is also used to tune a model (e.g. RMs and RF, see Subsection 2.2). To perform bootstrap validation, one repeatedly fits the model in bootstrap samples (of the same size as the original sample, but drawn with replacement) and evaluates the performance of the model on the original sample. The estimate of the likely performance of the model on new data is estimated by the average of all the bootstrap sample model estimates of performance computed on the original sample. This estimate is slightly biased downward (Hastie, Tibshirani, and Friedman 2008). In order to reduce the bias in the bootstrap estimate of model performance, Efron (1979, pages 247-252) suggested estimating the optimism in the model and then subtracting it from the apparent performance, derived from the original sample, to obtain a bias-corrected estimate of performance. This improved version of bootstrap validation was later reproposed by Harrell, Lee and Mark (1996) and Steyerberg (2009) and will be used in this project as the preferred internal validation method (refer to the paragraph below).

Internal validation can be used to recalibrate the model and obtain improved performance on new data. Thus, when presenting the model (step seven in Figure 1.1), the recalibrated model should be presented if calibration performance is not optimal. For example, validated estimates of calibration measures can be used to recalibrate a linear regression model $M(\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{b}$, with $\boldsymbol{X}$ being the matrix of explanatory variables included vector $\underline{1}$ for the intercept and $\boldsymbol{b}$ being the vector of coefficients, this way: $M_{recalibrated}(\boldsymbol{X}) = \alpha_{validated} + \beta_{validated}\boldsymbol{X}\boldsymbol{b}$. Thus, shrinkage of $\boldsymbol{b}$ will occur if $\beta_{validated} < 1$ and unshrinkage of $\boldsymbol{b}$ if $\beta_{validated} > 1$ (Steyerberg and Vergouwe 2014).

#### Validation in statistical learning

In statistical learning, the simplest and most widely used method for estimating prediction error is *cross-validation* (CV, Hastie, Tibshirani, and Friedman 2008). Let us call $\mathcal{T}_i = (\boldsymbol{x}_i, y_i), i = 1, \ldots, n$ the observations of the individual $i$, where $\boldsymbol{x}_i$ is the vector of inputs and $y_i$ is the outcome. Let $\mathcal{T} = \{(\boldsymbol{x}_1, y_2), \ldots, (\boldsymbol{x}_n, y_n)\}$ be an independent and identically distributed (i.i.d.) sample from the multidimensional distribution $F$. If we estimate the model $\hat{f}_{\mathcal{T}}$ from our data $\mathcal{T}$, then $\hat{\boldsymbol{y}} = \hat{f}_{\mathcal{T}}(\boldsymbol{x}_0)$ is the predicted value of $y$ at $\boldsymbol{x} = \boldsymbol{x}_0$. We call $L[y, \hat{f}(\boldsymbol{x})]$ the measure of error between the response $y$ and the prediction $\hat{f}(\boldsymbol{x})$. Therefore, the *expected prediction error* (also called *true error*, *generalization error* or *extra-sample error*) for $\hat{f}_{\mathcal{T}}(\boldsymbol{x}_0)$ is defined:

$$\text{Err}_{\mathcal{T},F} := E_{0,F}\left[L[y_0, \hat{f}_{\mathcal{T}}(\boldsymbol{x}_0)|\mathcal{T}]\right], \tag{1.4}$$

where the notation $E_{0,F}$ means the expectation over a new observation $(\boldsymbol{x}_0, y_0)$ from the population $F$ and $\mathcal{T}$ is fixed. Thus, $\text{Err}_{\mathcal{T},F}$ refers to the error for this specific training set $\mathcal{T}$, and therefore it is a conditional error.

CV directly estimates the *expected extra-sample error* $\text{Err}_F$ (Hastie, Tibshirani, and Friedman 2008), i.e. the average of $\text{Err}_{\mathcal{T},F}$ over training sets $\mathcal{T}$:

$$\text{Err}_F := E_{\mathcal{T},F} E_{0,F} \left[ L[y_0, \hat{f}_{\mathcal{T}}(\boldsymbol{x}_0) | \mathcal{T}] \right], \tag{1.5}$$

when the method $\hat{f}_{\mathcal{T}}(\boldsymbol{x})$ is applied to an independent test sample $\mathcal{T}$ from the joint distribution of $\boldsymbol{x}$ and $y$.

CV is preferred as a validation and model selection technique for its simplicity and for the fact that the training sets do not overlap with the test set, as on the contrary it happens with the traditional bootstrap technique. However, the training sets are not independent samples as they will have at least $k - 2$ folds of observations in common and also the test sets (left-out folds) come from the same data. In contrast, when using the bootstrap, the different bootstrap samples are independently and randomly drawn from the original sample. As a consequence, it is known that the CV estimate of the extra-sample error is biased upward as it is the estimate of the error for a smaller sample size, i.e the sample size of $k - 1$ folds (Borra and Di Ciaccio 2010, Hastie, Tibshirani, and Friedman 2008). Also its variance is larger than the bootstrap estimate. To some extent, the CV error depends on the initial random split. As a result, to avoid this and to reduce the variance, the process is repeated more times (for example 20 times for $k = 10$, Harrell 2001) with different random partitions (repeated CV by Burmann 1989). However, repeating the CV procedures many times can be computationally expensive and time consuming. On the other hand, the traditional bootstrap procedure computes slightly more biased estimates of the error because of the overlapping of training set (bootstrap sample) and test set (original dataset), but it yields smaller variance with a reasonably small number of resampling steps: 100 (number of bootstrap samples, Harrell 2001, Smith et al. 2014). The bias in the bootstrap estimate of error is reduced with the *.632+ bootstrap* estimator (Efron and Tibshirani 1997). However, this change in the bootstrap was supported by a heuristic argument instead of a theoretical justification (Arlot 2010), even though empirical studies show the improvement over the traditional bootstrap (Efron and Tibshirani 1997).

The improved version of bootstrap validation (see paragraph above, Efron 1979,Harrell, Lee, and Mark 1996 and see Subsection 2.1.3) and the CV validation methods can both estimate optimism in any model performance measures (Steyerberg 2009) in an almost unbiased

way (to note that in statistical learning, only the optimism in the extra-sample error is usually estimated and not in the calibration measures as recalibration is not popular when models are biased estimators). However, bootstrapping is the fastest method of performing optimism correction, as a smaller number of resampling steps are needed compared to CV in order to have stable estimates of performance.

### 1.2.2 Missing data

A problem for the application of statistical learning models like regularised regression methods in psychiatric studies is **missing data**. People not attending treatment visits and follow-up, or refusing to answer to items in questionnaires, or the researchers' error in measuring the variables may lead to substantial amounts of missing data in baseline variables and outcomes.

Missing data can be defined as data meant to be collected for studying a specific problem but were not (although data can also be missing by design, especially in secondary analysis). They are common, but they are often inadequately handled in both observational and experimental research (Wood, White, and Thompson 2004, Chan and D. Altman 2005 and Sterne et al. 2009). Typically, in prediction modelling a big amount of baseline data is used, as sometimes identification of predictors is the aim. This results in a larger quantity of missing data compared to the analysis of primary outcomes in RCTs, where generally only the outcome variable measured at baseline is needed.

In explanatory or inferential statistics missing data are well studied (Rubin 1976, Carpenter and Kenward 2007). The process by which observations become missing is called the *missingness mechanism*. However, the missingness mechanism is usually unknown and the data alone or the missingness pattern or its relationship to the observations cannot identify such a mechanism. Therefore, assumptions are usually made on the missingness mechanism in order to be able to analyse the data to draw sensible inferences. Rubin (1976) defined three classes of missingness mechanisms in a pure likelihood/Bayesian way, but here I will use a frequentist definition (Carpenter and Kenward 2013) and the notation is given below. Given a sample of $n$ individuals from an infinite population, let $\boldsymbol{y_i} = (y_{i,1}, y_{i,2}, \ldots, y_{i,p})^T$ represent the $p$ measurements of the variables intended to be collected for the $i^{th}$ individual, $i = 1, \ldots, n$. Let $\boldsymbol{y_{i,O}}$ be the subset of $p$ variables that are observed for the individual $i$ and $\boldsymbol{y_{i,M}}$ be the subset that are missing. Then let $R_{i,j}$ be the binary indicator of missingness for each unit $i$ and variable $j$: if $y_{i,j}$ is observed, then $R_{i,j} = 1$ and if $y_{i,j}$ is missing, then $R_{i,j} = 0$. Let us define $\boldsymbol{R_i} = (R_{i,1}, R_{i,2}, \ldots, R_{i,p})^T$. The three classes of missingness mechanism are:

- **Missing completely at random (MCAR)**: The data are MCAR if the missingness mechanism neither depends on covariates relevant to the analysis, nor does it depend on the unseen data. In other words, the probability of a value being missing is unrelated to the observed and unobserved data for that individual:

$$P(\boldsymbol{R}_i|\boldsymbol{y}_i) = P(\boldsymbol{R}_i)$$

  Here, the incomplete population is a random sample of the complete population, then the subjects with missing data are representative of the population. For example: missing observations because a page of the questionnaire was missing, or because of a data processing error, or due to a change in the data collection procedure. In this case analysing only those individuals with observed data gives valid and unbiased results, even though the estimates will be less precise than when in presence of a complete data set. However, MCAR data are not always plausible.

- **Missing at random (MAR)**: The data are MAR if the missingness mechanism does not depend on the unobserved data conditional on the observed data:

$$P(\boldsymbol{R}_i|\boldsymbol{y}_i) = P(\boldsymbol{R}_i|\boldsymbol{y}_{i,O})$$

  The probability of missingness may depend marginally on the unobserved data but it is independent of the missing data when we condition on the observed data. For example, if the probability of a missing observation depends on an earlier observation, after accounting for the earlier observation, the probability of observing the missing observation is independent of its value. Therefore, if only the subjects with complete data were analysed, we would get invalid results because they will be biased as the fully observed subset of data will not be representative. Also, there will be loss of information since we would have thrown away information on cases with even one missing observation. To obtain valid estimates, the variables predictive of non-response need to be included in the analysis (e.g. covariates in a regression). In addition, simple summary statistics are invalid as estimates of population parameters under MAR. It is important to note that it is not possible to assess any residual dependence between missingness mechanism and the missing variable, i.e. the assumption of MAR cannot be tested.

- **Missing not at random (MNAR)**: The data are MNAR if the chance of missingness de-

pends on the unseen data, even after conditioning on all the observed data:

$$P(\boldsymbol{R}_i|\boldsymbol{y}_i) \neq P(\boldsymbol{R}_i|\boldsymbol{y}_{i,O})$$

It is difficult to analyse MNAR data because under the MNAR assumption conditional distributions of partially observed variables are not the same in individuals with and without observed data (as they are under MAR). Consequently, under MNAR one will need to model both the response of interest and the missingness mechanism through sensitivity analyses and may expect quite different conclusions from different models.

In statistics, Multiple Imputation (MI, Rubin 1981) is a popular method to impute (that is fill in) missing data. MI consists of creating $m$ different imputed datasets from the original dataset with missing observations. Then each of the $m$ completed datasets are analysed and the $m$ analysis results are pulled into one valid result through Rubin's Rules (Rubin 1981). An extension of MI is Multivariate Imputation using Chained Equations (MICE, Van Buuren and Oudshoorn 2000, see Subsection 2.1) which allows the imputation of data in the presence of missing multiple outcome values (e.g. hierarchical or longitudinal data) and missing predictors. Inferences from analyses with multiply imputed data are only valid when data are MAR. However, any suitable additional variable measurements (i.e. auxiliary variables predictive of the missing values, Hippel and Lynch 2013), included in the imputation model, provide more observed data to condition on and thus make the MAR assumption more plausible (Carpenter and Kenward 2007, Ibrahim, Lipsitz, and Horton 2001 and Rubin, Stern, and Hehovar 1995). However, if too many auxiliary variables are included in the imputation model of regression analyses with missing predictors for example, parameter estimates result biased downward and less precise, mainly when the correlations between variables are low and the sample size is small Hardt, Herke, and Leonhart 2012. This would also affect prediction accuracy of the model.

#### Missing data in statistical learning

When there are missing data in the dataset meant for the analysis, the usual approach in statistical learning is to analyse only cases with complete observations or impute the missing observations with plausible values through specific techniques depending on the missingness mechanism. There are three approaches for treating missing data:

1. if the data are MCAR, one can run a *complete records analysis*, i.e. discard observations with any missing values, when the percentage of missing data is low (the validity of this method is stated above),

2. impute the missing values (commonly through mean substitution) before training the model but this will add uncertainty to the analysis: if resampling is used to estimate validated performance and imputation is not incorporated in the resampling process, estimates of model performance will be biased (Kuhn and K. Johnson 2013),

3. or impute the missing data in the model training phase: the information in the training set is used to predict the test set missing data. This strategy incorporates the uncertainty due to imputation in the analysis and is preferred.

The following statistical learning methods deal effectively with missing data using approach 2, also when the percentage of missing data is high: (non-parametric) CART (Breiman, JH Friedman, et al. 1984) and (parametric) MARS (when the data are MAR, Hastie, Tibshirani, and Friedman 2008). CART fills in missing data through surrogate splits. MARS will automatically impute missing data by estimating the joint probability distribution of the data $P(\boldsymbol{X}_i, \boldsymbol{y}_i)$ and sampling $\boldsymbol{X}_i, \boldsymbol{y}_i$ pairs (Bishop 2007). While CART based imputation methods are able to impute high dimensional datasets with high accuracy without making any assumption on the data, MARS in datasets with a large number of variables require a number of samples that increases exponentially with the variables number for accurate estimation, if there are no appropriate model assumptions. Also Random Forest (RF) can handle mixed type of missing data and Tang and Ishwaran (2017) lately wrote a comprehensive review of the different RF missing data algorithms. They included the original RF proximity algorithm proposed by Breiman (2003) and implemented in the `randomForest` R-package (Liaw and Wiener 2002b), the 'on-the-fly-imputation' (OTFI) algorithms implemented in the `randomSurvivalForest` R-package (Ishwaran, Kogalur, et al. 2008), which allow data to be imputed while simultaneously growing a survival tree, and the algorithm MissForest (implemented in the R package `missForest`, Stekhoven and Buhlmann 2012 and see Subsection 2.1.2) that takes a different approach by remodelling the missing data problem as a prediction problem. The first two algorithms have been unified within the `randomForestSRC` R-package to include not only survival, but classification and regression settings (Ishwaran and Kogalur 2016). By comparing the RF missing data algorithms with a simulation study, Tang and Ishwaran (2017) found that the imputation performance of all RF procedures improved (i.e. the imputation error decreased) with increasing correlation of features. Moreover, MissForest had the least imputation error when there was high correlation between variables, even though it had the longest computational time. In the simulation study, all RF algorithms on average performed well with MCAR and MAR data, but performance in NMAR data was generally poor unless correlation was high (>0.8).

Other techniques used in statistical learning to impute missing data are: (non-parametric) K-nearest neighbours (KNN, Jönsson and Wohlin 2004, Liao et al. 2014) and the (parametric) Expectation Maximisation (EM, Dempster, Laird, and Rubin 1977) algorithm. The KNN method imputes a missing value of an individual using values calculated from the $k$ individuals with observations close to the ones of the individual with the missing value. The replacement value is the mode of the nearest neighbours in case of discrete data or the mean for continuous data. However, in this local method the nearest neighbours will not be close to the target individual in high dimensional input spaces and this can result in high variance (Hastie, Tibshirani, and Friedman 2008). KNN is somewhat related to RF (Tang and Ishwaran 2017). Both methods are a type of nearest neighbour method, although RF is more adaptive than KNN and in fact can be more accurately described as an adaptive nearest neighbour method. The EM algorithm deals with missing data by marginalising the joint distribution of the observed variables over the distribution of the unobserved variables and it requires MAR data for the imputation to be reliable. However, the EM algorithm is slow, computationally expensive with high dimensional data and often does not converge if the percentage of missing data is high.

In statistical learning, a common approach is to treat the imputed missing values as observed, but this ignores the uncertainty due to imputation that in turn will add uncertainty into the predictive accuracy of the models used on imputed data. One solution to this is measuring this additional uncertainty by imputing multiple times and hence creating many different training sets (like MI). Therefore, the predictive model for $\boldsymbol{y}$ can be fit to each completed training set and the variation in the performance across training sets can be assessed through estimating the mean performance (Hastie, Tibshirani, and Friedman 2008).

The third approach mentioned above to treat missing data in prediction modelling (i.e. incorporating the imputation in the model training process) is only possible when the missing data imputation technique can develop an imputation model on the observed training set, able to impute the test data by using the training data only. This will properly validate the model incorporating the missing data technique. KNN and bagged-tree imputation (Kuhn and K. Johnson 2013) can do likewise, but the dataset with missing values will need to contain a sufficiently large number of complete records (i.e. a study participant for whom all the variables are measured without missing values), on which to build the imputation model. However, sometimes the data only contain a small number or no complete records at all. Therefore, methods like MICE (see above), RF by proximities or MissForest can only be applied independently to training and test data by introducing some bias in the validated performance measures. Internal validation techniques like Efron's bootstrap as for Harrell et al. (1996) aim to correct validation bias.

In many clinical trial datasets, including the data available for this project, there are high percentage of missing data and no complete records. Therefore, parametric MICE and the best RF imputation technique (MissForest) will be considered (see Subsection 2.1.2).

**Combining missing data imputation and regularised regression methods**

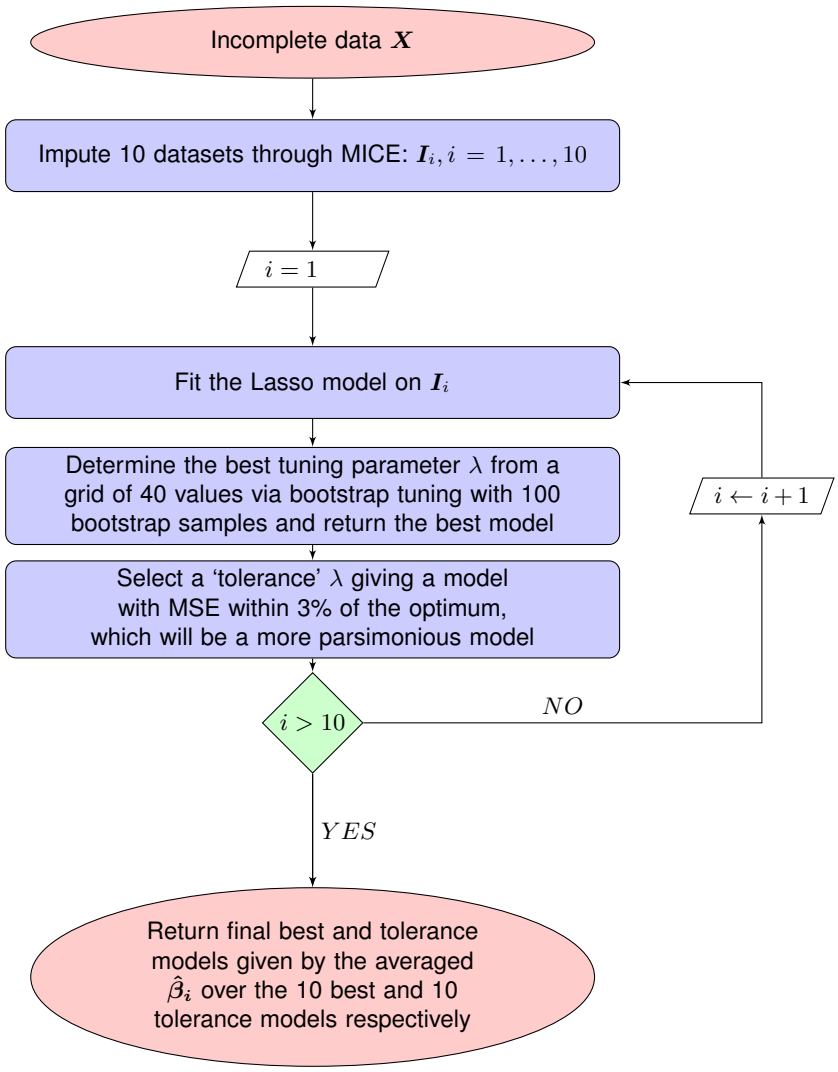
In 2014 (i.e. when this PhD project started), to my knowledge, three main publications proposed regularised regression methods combined with imputation methods:

1. *Multiple imputation-least absolute shrinkage and selection operator (MI-LASSO*) method by Chen and Wang (2013) combining MICE and the Group-Lasso (Yuan and Lin 2011),
2. *MICE combined with the Lasso* by Musoro et al. (2014),
3. *Imputed-LASSO* by Lu and Petkova (2014) combining Random Forest (RF) imputation by proximities and the Lasso.

In their method MI-Lasso, **Chen and Wang (2013)** applied the Group-Lasso to the multiply imputed datasets jointly, not to each imputed dataset separately, i.e. they used the large stacked dataset formed by the different multiply imputed datasets altogether. This allowed them to select the variables consistently across the imputed datasets by minimizing the sum of squares penalized through the Group-Lasso penalty. Therefore, the estimated coefficients for the group, formed by the same variable across multiply imputed data, would either all be exactly 0 or all be different from 0. Then the final covariate coefficients were estimated by applying Rubin Rules (RR, Rubin 1981) on the groups of coefficients selected by the Group-Lasso. The authors also presented a simulation study (linear regression data-generating model for the outcome, multivariate normal distribution, continuous or dichotomous covariates, compound symmetry or first-order autoregressive covariance structure, $p = 20$ or $40$, $n = 100$ or $200$) comparing Lasso and the traditional stepwise selection method in absence of missing data, and the MI-LASSO and RR-stepwise selection method (Wood, White, and Royston 2008) in presence of missing data (MCAR or MAR, 60% or 35% complete cases). Simulation results showed that MI-LASSO could identify the true predictors in a linear regression model similarly to Lasso when the data were complete and the selected models had similar $MSE$s. When missing data were present, MI-LASSO was superior to Lasso complete records analysis and to the stepwise regression methods. However, the authors did not validate their method internally or externally. Furthermore, the disadvantage of the Group-Lasso variable selection performance (as for the Lasso) is its vulnerability to high correlation between variables: groups of noise variables are more likely to be selected (Yunus 2017).

**Musoro et al. 2014** combined the Lasso with MICE and validated the method internally via Efron's bootstrap validation (Harrell, Lee, and Mark 1996 and Steyerberg 2009). The authors aimed to quantify optimism in the predictive performance of Lasso and investigate how internal validation should be applied in the presence of multiply imputed data. Their proposed model is illustrated in Figure 1.3. They ran a Lasso model with bootstrap tuning on each of 10 imputed dataset and retained the models corresponding to the best penalty $\lambda$ (best model with minimum MSE) and to a tolerance penalty $\lambda$ giving an MSE which was within 3% of the optimum (3% tolerance model, more parsimonious in the variable selection).

Figure 1.3: Musoro et al (2014) model.




To validate the model prediction accuracy through bootstrap validation (Harrell, Lee, and Mark 1996, see the Subsection 2.1.3) with 100 bootstrap samples, the authors analysed four approaches in order to handle the multiply imputed datasets in the bootstrap resampling:

1. in a bootstrap run the same subjects across the imputed data sets were selected, so that the bootstrap samples differed only by the imputed data values,
2. records selected in a bootstrap run could differ over the imputed datasets,
3. only one of the imputed data sets was selected and resampling procedures were performed as in the case of absence of missing data,
4. the MI procedure was incorporated in the validation, i.e. the proposed model was run on each randomly drawn bootstrap sample.

The fourth approach is the one which is theoretically valid, as it includes the MI uncertainty in the validation process and this was confirmed in the simulation study results that the authors ran (see below). Performance was assessed through validated discrimination and calibration, i.e. by looking at bootstrap corrected MSE and calibration slopes. When a model was miscalibrated (i.e. the agreement between predicted and observed outcome was poor), it was recalibrated by using the bootstrap-corrected calibration-in-the large and the calibration slope.

The authors ran a simulation study with 20 covariates (10 continuous and 10 binary variables) drawn from a multivariate standard normal distribution, with a sparse correlation matrix. The outcome was predicted by 5 continuous and 5 binary variables according to a linear regression model. Eight variables contained MCAR missing values in the percentage of 20% and 50%, but the outcome was complete. Such datasets were simulated 1000 times with 250 and 1000 observations. Two further independent datasets (with 250 and 1000 observations respectively) used for external validation were generated. Lasso was first fitted on the complete data in order to estimate the apparent MSE and obtain an estimate of the external MSE by fitting the trained model on the simulated external data. Following this, the expected optimism was estimated via bootstrap resampling as for Harrell, Lee and Mark (1996) and via the four approaches with and without missing data through MICE-Lasso and Lasso respectively.

In the case of complete data and 250 observations, the best Lasso model almost always selected all the true predictors, but also retained a large number of irrelevant covariates (with the selection frequency ranging from about 45% to 55%) as expected (Fan and Lv 2009, see Subsection 2.1.1). This was even larger (66% to 75%) in the case of missing data because covariates were counted if they were included in at least one of the imputed datasets. A better variable selection was performed by the tolerance models (25% to 47% with missing data). With 1000 observations there was much lower selection of noise covariates (3% to 25% with missing data). Both missing data settings had optimistic apparent MSEs.

The estimated optimism of approaches 1 to 3 were optimistic and only approach 4 showed an optimism similar to the optimism obtained comparing apparent and external validation (slightly biased upward). This confirmed the superiority of the fourth approach. The calibration estimates showed there was over-shrinkage of coefficients.

Musoro et al. (2014) compared their method to MI-LASSO (Chen and Wang 2013) and the performance of the tolerance model from Musoro et al. (2014) was similar to the performance of MI-LASSO.

**Lu and Petkova (2014)** combined RF imputation with the logistic Lasso in a method they called Imputed-LASSO. The RF imputation prior to applying the Lasso was a modification of the original RF imputation by proximities as for Breiman (2003). The imputation was iterative (four iterations): in the initial iteration, median imputation was done and a RF model was built on the imputed data, then proximities (measures of the similarity between two subjects given by the proportion of times they end up in the same final leaf) for the imputed data were calculated and new imputations were obtained from the average or vote weighted by these proximities. In the remaining iterations, a new forest was built on the new imputed data from the previous iteration, and proximities and imputations were updated accordingly. Only *one imputed dataset* was created.

A complex simulation study was run in order to compare different variable selection methods (including CART, for which the variables that appear in the tree were considered as selected; RF, for which only the top 10 variables were considered as selected; and Elasticnet) with the Imputed-Lasso method in presence and absence of missing data. The simulations covered many different scenarios: complete data and equal prevalence of cases and non-cases, missing observations (5% MCAR giving 50% complete cases) and equal prevalence, unequal prevalence of cases and non-cases. The noise and true predictors (10 true predictors and 50 noise variables for a sample size of 400) were generated to have the following six situations: independent predictors, correlated (true-noise or true-true) predictors, predictor interactions omitted or included and presence of unobserved true predictors. The variables were described by the authors as being categorical (3 levels), but in the model specifications the dummy coded levels did not appear and results were reported as if the variables were treated as continuous. Results showed that in absence of missing data and equal prevalence, the LASSO and Elasticnet performed better than CART and RF (top ten variables) in variable selection when there was correlation between true and noise predictors, but in the case of correlated true predictors Lasso performed less well than Elasticnet and RF outperformed the other methods. When there were missing observations and equal prevalence of cases, CART showed the same results as

in absence of missing data revealing robustness to sample size decrease with MCAR data. RF performed well when there was equal probability of missingness between variables (chapter 'Tree-based methods' in Cutler et al. 2009), but for the unequal probability of missingness scenario RF tended to select more often the noise variables with large probability of missingness as expected (Strobl, AL Boulesteix, and Zeileis 2007). Imputed-LASSO excelled in selecting true predictors with both equal and unequal probabilities and chose fewer false positives than regular LASSO, especially when the interactions were included in the model. Imputed-LASSO selection was not affected, even though after RF imputation the correlation between variables with more missing data and the outcome variable increased more than the correlation between those with less missingness. However, in the case of correlations between true predictors, the increment in correlation between predictors and response due to RF imputation caused the Imputed-LASSO to be much more likely to select the true variables with larger missingness only. This confirmed the inconsistency in selecting strongly correlated true variables (correlation = 0.8) that characterizes LASSO (Zou and Hastie 2005). Imputed-LASSO also performed better than other methods in terms of prediction accuracy, apart from the case of correlated true predictors with equal missingness probability, where RF achieved best prediction. For the scenarios with unequal prevalence of cases and non-cases, again Imputed-LASSO outperformed the methods above, except for the case in which half of the true predictors were not included in the model. This suggests that if data are unbalanced and regularised regression methods select only a few predictors, then this might be because of a large proportion of missing observed predictors. Finally, regarding prediction accuracy, Elasticnet and LASSO outperformed the other methods when data were complete, and in the case of incomplete data, Imputed-LASSO was superior. Split-sample validation was used to internally validate the methods.

In order to assess which method combining missing data imputation techniques and statistical learning models performs best in prediction accuracy and at the same time allows interpretability, I will compare Musoro et al.'s (Musoro et al. 2014) MICE-Lasso with a technique similar to the one proposed by Lu and Petkova (2014) using Efron's bootstrap internal validation as for Harrell, Lee and Mark (1996). While Lu and Petkova (2014) used a RF imputation method by proximities followed by the Lasso, I will use the more accurate RF imputation method MissForest (Stekhoven and Buhlmann 2012, Tang and Ishwaran 2017), which uses iterative RF predictions to fill in the missing data, before applying the Lasso. Then I will also consider the case of categorical variables and the ratio 1:2 between number of model covariates and sample size. Furthermore, I will compare MICE-Lasso and MissForest-Lasso with a RF model after MissForest imputation through a simulation study. Group-Lasso combined with MICE

(MI-Lasso, Chen and Wang 2013) performed slightly better in variable selection compared to MICE-Lasso, according to the reported simulations of Musoro et al. (2014). However, I decided to use the Lasso penalty instead of Group-Lasso because the model that the PhD project aims to develop needs to be accessible by non-statisticians and clinicians, and Lasso is easier to understand and more popular than Group-Lasso.

Recently, two more methods combining regularised regression and data imputation, not known at the time I chose the methods, have been introduced:

- MI-based weighted Elasticnet or MI-WENet (Wan et al. 2015): Elasticnet is run on stacked multiply imputed (through MICE) data with weights accounting for the proportion of the observed information for each observation,
- Multiple Imputation Random Lasso or MIRL (Liu et al. 2016): Random Lasso has an improved performance compared to the Lasso when the correlation among variables is high. MIRL has been shown to deal with high proportions of missing data (50%) efficiently.

These two methods were not compared with the above three methods.

### 1.2.3 Dimension reduction of multiple outcomes

In psychiatric research, disease complexity is often not adequately characterised by a single outcome. This is the case of **commensurate outcomes**, i.e. multiple observed outcomes (Teixeira-Pinto et al. 2009) measuring the same underlying construct using the same scale. The cognitive outcomes of the clinical data available for this PhD project are commensurate outcomes. In particular, there is an interest in executive function, processing speed and memory outcomes which are descriptions of different aspects of cognition. Most statistical learning methods can handle only one outcome unlike traditional statistics which offers a variety of modelling approaches, such as joint modelling and generalized linear mixed models. Multivariate regularised generalised linear mixed models can deal with commensurate outcomes, but such methods had only just been developed when I started the project, with slow and not well debugged algorithms (see Subsection 1.1.4). Therefore, I will only use univariate regularised regression methods in the PhD and will then need to summarize the different observed cognitive outcomes in one variable. A common procedure to summarize outcomes measuring an unobserved (latent) construct is Factor Analysis (FA).

Factor analysis is a statistical technique which postulates that the correlation of the observed variables is explained by a smaller number of underlying unobserved variables and tries to identify and test the measurement model (Lattin, Carroll, and P. Green 2003). Exploratory

FA (EFA), a technique within FA, aims to determine the underlying relationships between measured variables. It is first used when no a priori hypothesis about factors or patterns of the multiple observed outcomes is assumed. In a second step, confirmatory FA (CFA), another special form of FA, is used to test whether measures of a construct are consistent with the EFA suggestions, also using an independent data set. FA can provide a valid and reliable aggregate measure for the latent summary variable (the factor scores). Models requiring a single outcome like regularised regression methods (Lasso and Elasticnet) can then be applied.

Therefore, I will apply FA to estimate factor scores of a summary latent measure for the multiple outcomes. The factor scores will then be used as a main clinical outcome measure for the prediction model I aim to develop.

### 1.2.4 Summary

People with SCZ experience cognitive difficulties and these are associated with poor functional outcomes (Rajji, Miranda, and Mulsant 2014). There is evidence for the effectiveness of CRT treatment in reducing the cognitive problems of SCZ (Wykes, Huddy, et al. 2011). Identifying CRT predictors of differential response using moderation analysis of individual participant data from different RCTs would help to minimise treatment response heterogeneity of outcomes, as well as contributing to personalised treatment and better prognosis. However, identifying moderators of CRT involves analysing high dimensional data with large percentages of missing data in predictors and outcomes. Therefore, the implementation of suitable missing data imputation techniques such as MICE (Van Buuren and Oudshoorn 2000) and MissForest (Stekhoven and Buhlmann 2012), combined with statistical learning methods such as regularised regression methods (Lasso, Elasticnet, Hastie, Tibshirani, and Friedman 2008) and RF (Breiman 2001) is needed. Regularised regression methods are good for ease of interpretation, while RF is optimal in prediction accuracy.

A simulation study will assess the most suitable combined method as the one showing the best trade-off between accuracy and interpretability. The chosen method will be then applied to the clinical data. Because of the univariate nature of regularised regression methods and because there is an interest in analysing multiple outcomes simultaneously, FA will be used to summarise commensurate outcomes with one latent outcome. A precision medicine prediction model, with the summary measure as dependent variable, will be developed.

## 1.3 Thesis aims and objectives

The primary project aims consist of:

- developing a robust prediction model for precision medicine using computer intensive statistical learning methods (Hastie, Tibshirani, and Friedman 2008) able to deal with large percentages of missing data in the predictors, and lower percentages of missing data in the outcome. This will allow the analysis of large incomplete psychiatric data, by maximising the quantity of information they provide;
- identifying moderators of cognitive remediation therapy (CRT, Wykes, Brammer, et al. 2002) in people with SCZ. A precision medicine prediction model would allow clinicians to tailor a treatment to an individual patient according to their characteristics, and to understand the mechanism responsible for differential treatment responses.

These are the statistical aims of the project:

1. To address common key problems of psychiatric studies simultaneously such as:
    - missing data
    - variable selection or measurement of variable importance in the model
    - overfitting
    - multicollinearity
    - analysis of multiple studies' individual data using statistical learning
    - factor analysis of commensurate outcomes
    - longitudinal invariance of a latent factor

    Please see Sections 1.1 and 1.2 for definitions and explanations;
2. To explore the best trade-off between prediction accuracy and interpretability in the choice of the most suitable statistical learning model.

## 1.4 Thesis structure

The thesis is formed of 2 main chapters:

### 1.4.1 Simulation study

In Chapter 2, I will compare the following prediction models combining missing data imputation techniques (MICE and MissForest) with statistical learning methods such as regularised regression methods (Lasso and Elasticnet) and Random Forests through a simulation study. Different missing data scenarios, correlation between covariates and sample sizes will be analysed. The model returning the best compromise between validated prediction accuracy and variable selection performance will be then used to develop the precision prediction model for CRT in Chapter 3.

### 1.4.2 Prediction model development

In Chapter 3, I will present the multiple randomised controlled trials' individual participant data used in the project with respective information and summary statistics. Next, I will run a factor analysis of the cognitive outcomes in order to find a summarising latent factor, which will be used as a dependent variable in the development of a precision medicine model for CRT. Then, I will build another precision medicine model with one of the cognitive outcomes as the dependent variable; the particular cognitive outcome was chosen because of its clinical importance and popularity in the literature. This model will allow me to study the effect of imputing missing dependent variable values on prediction accuracy and feature variable selection. Finally, the performances of the developed models will be compared and results discussed in light of the existing literature.

# Chapter 2

# Prediction modelling combining statistical learning with missing data imputation: a simulation study

## 2.1 Introduction

In the previous chapter I identified two modelling approaches for clinical prediction models, which are a compromise of showing good prediction accuracy, ability of variable selection and clinical interpretability, namely regularized regression (the Least absolute shrinkage and selection operator or Lasso by Tibshirani 1996 and Elasticnet by Zou and Hastie 2005) and Random Forests (Breiman 2001). Both modelling approaches are known to handle high-dimensional data sets.

Lasso and Elasticnet were proposed because of their variable selection property and interpretability, while Random Forest was chosen because of its generally good prediction accuracy, ability to model more complex relationships and variable importance measures (James et al. 2013). Regularised regression methods also allow to handle missing data appropriately when combined with imputation methods (MICE in Chen and Wang 2013 and Random Forests imputation by proximities in Lu and Petkova 2014). However, it is not known how well these combined models perform in selecting the correct variables (or estimating the correct variable importance) when applied to complex data sets, e.g. including categorical variables with more than two levels and datasets with the number of covariates in the model (also comprising interactions terms) close to the sample size. Similarly, it is not known how good the discrimination and calibration abilities of the models to predict unseen cases is. Because for clinical datasets

the truth is not known, Monte Carlo (MC) simulations (Burton et al. 2006) were used to assess the performance of such methodologies. By simulating data from known probability distributions as inputs to model uncertainty and known relationships between independent and dependent variables with different missing data mechanisms and correlation matrices, I can evaluate the characteristics and behaviour of complex processes in the comparison of the expected and observed behaviour of a modelling technique.

In this chapter, I will assess the performance of the above three statistical learning techniques (Lasso, Elasticnet and Random Forests) combined together with two missing data handling procedures (Multivariate imputation using chained equation (MICE) by Van Buuren and Oudshoorn 2000 and another iterative method (MissForest) based on random forests predictions by Stekhoven and Buhlmann 2012) using MC simulations in terms of

- variable selection
- prediction accuracy
- ability to handle missing data
- clinical interpretability and usefulness.

Namely, the proposed methods evaluated in this simulation study will be the following combinations of the mentioned methods:

- MICE-Lasso (Musoro et al. 2014) and MICE-Elasticnet: Lasso (or Elasticnet) was run on the imputed datasets and averages of the estimated coefficients across imputed datasets constituted the final model coefficients
- MissForest-Lasso and MissForest-Elasticnet: single MissForest imputation was followed by Lasso (or Elasticnet)
- MissForest-RF and MissForest-Conditional RF: single MissForest imputation was followed by RF (or Conditional RF)

MICE-RF will not be considered in the simulation study because the main purpose of using RF was to compare its usually very good prediction accuracy with the less good accuracy of the more interpretable regularised regression methods (Lasso and Elasticnet). Therefore, only RF combined with a RF imputation method was of interest to this aim.

The method showing the best trade-off between prediction accuracy and model parsimony to allow clinical interpretability will be chosen for the analyses of the cognitive remediation therapy RCT data sets.

In the following sections I will first describe the statistical learning methods and missing data procedures and how the performance was assessed using bootstrap validation as for Harrell, Lee and Mark (1996). Next, I will explain the concept of MC simulations, how they are performed in classical inferential statistics and how they need to be adapted to assess prediction models using statistical learning methods. Then, I will justify the choice of the main simulation parameters to estimate, introduce the different simulation scenarios, and present some hypotheses of what I expect from the simulation results, according to the knowledge to date. Finally, I will explain how simulations were conducted and report the results.

### 2.1.1 Statistical learning methods for prediction modelling

#### Random forests and Conditional Random Forests

Random Forests (RF, Breiman 2001) is a statistical learning non-parametric method based on Classification and Regression Trees (CART, Breiman, JH Friedman, et al. 1984).

**CART** CART are non-parametric prediction models which perform *recursive binary splitting* to partition the covariates space into disjoint regions $R_1, \ldots, R_G$. Then, every test observation that falls into a particular region is predicted with the mean (regression) or the mode (classification) of the outcome for the training observations in that region. Recursive binary splitting finds the best split at each step according to a performance criterion: minimising the residual sum of squares (RSS) for continuous outcomes and the classification error rate or the Gini index or the cross-entropy for binary outcomes. In the case of regression trees, at each step all predictors $X_1, \ldots, X_p$ are considered together with all possible values of cutpoints $s^j_{i_j}$ for each of the predictors ($j$ referring to the predictor $X_j, j = 1, \ldots, p$, and $i_j$ referring to the different values taken by predictor $X_j$, with $i_j = 1, \ldots, n_j$). Then, the predictor and respective cutpoint that minimise the RSS of the resulting tree are chosen. Thus, if at the first step $X_k$ and $s^k_{l_k} =: s$ are the chosen predictor with respective cutpoint, $R_1(k, s) := \{X|X_k < s\}$ and $R_2(k, s) := \{X|X_k \geq s\}$ are the two half-hyperplanes in which the space is divided first. This also means that the RSS for the tree generated by this first split is the minimum of the possible RSSs obtained from other combinations of predictor and cutpoint:

$$\text{RSS}(k, s) = \sum_{i:x_i \in R_1(k,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:x_i \in R_2(k,s)} (y_i - \hat{y}_{R_2})^2 \tag{2.1}$$

where $\hat{y}_{R_g}, g = 1, 2$ are the mean responses for the training observations within the $g$-th box (region), $g = 1, 2$ respectively. Then, the same procedure is repeated recursively in each of the regions created in the previous step and so on until a stopping criterion is met, e.g. until there is a maximum of four observations in every region. The predictions for the test observations in a particular region are the mean responses for the training observations belonging to that region.

Since $T_0$, the large tree constructed this way, may predict well the training observations, but could overfit the test data for its complexity, it is practice to do some *tree pruning* to generate a simpler subtree with lower variance but some bias. To do that, a cost complexity pruning is applied to $T_0$ in order to avoid computing all the subtrees test errors using computationally costly cross-validation procedures: for each value of a sequence of a tuning parameter $\alpha > 0$ there is a subtree $T \subset T_0$ that minimises

$$\sum_{m=1}^{|T|} \sum_{i:x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha |T| \tag{2.2}$$

where $|T|$ indicates the number of the terminal nodes of $T$ and $R_m$ is the box that corresponds to the $m$th terminal node. Then, through cross-validation the subtree is tuned on $\alpha$.

**Random Forest** RF is given by the combination of a large number of trees and the RF algorithm for continuous outcomes (similar for categorical outcomes) is the following:

1. Draw a bootstrap sample from the training data (sample drawn with replacement of the same size as the training data).

2. Develop an unpruned tree on the bootstrap dataset as follows:

   (a) Draw a random sample of the predictors of size $m$ without replacement ($m = p/3$ is adviced if the outcome is continuous, $m = \sqrt{p}$ if the outcome is binary, or $m$ can be treated as a tuning parameter to minimise the RF error and chosen through resampling methods).

   (b) Construct the first recursive binary split of the data.

   (c) Repeat step 2a for each subsequent split until a stopping criterion which for example could be that the tree needs to have maximum four observations per region (terminal node). Compute each region mean.

   (d) Use the developed tree to predict the out-of-bag (OOB) datasets (data in the training

sample not taken in the bootstrap samples, used as test sets) and compute the OOB mean squared error between predicted and observed outcome for the tree.

3. Repeat steps 1-2 a large number of times (e.g. 500).
4. Average the trees OOB mean squared errors to obtain the OOB error for RF.
5. Average the trees predictions to obtain the RF predictions.

Thus, the accuracy of the prediction can be estimated by testing each tree on the out-of-bag (OOB) sample, which comprises the observations in the original sample not included in the bootstrap sample where the individual tree is trained. The *OOB error* is the mean squared error (MSE, for continuous outcomes) or misclassification error (for categorical outcomes) averaged over the errors obtained testing each tree on the OOB samples. RF often yields a favourable error rate, it is known to perform very well with high dimensional data with large number of variables compared to the number of observations, in presence of complex interactions and non-linear data structures and it can handle missing data (see the Subsection 1.2.2). It can assess *variable importance* determined by the mean decrease in OOB accuracy (or mean decrease in node impurity from splitting on the variable for binary outcomes) or by a permutation test (which is implemented in the R package '`randomForest`' by Breiman, Cutler, et al. 2006). The permutation test works as follows: for each tree the OOB error is recorded, then the same is done after permuting each predictor by breaking the link with the outcome and any interacting variables and the differences between the two accuracies are then averaged over all trees and normalized by the standard error.

**RF with conditional inference trees** The above described RF variable importance measures are biased when variables have different scales of measurement or different number of categories: RF will prefer categorical variables with more levels and the continuous variables to the categorical variables with less levels (Strobl, AL Boulesteix, and Zeileis 2007). For example, in the case of different level of correlation among predictors, the described permutation test, which is unconditional, considers that each variable is independent of the response as well as of all other predictors. Since the correlated predictors are obviously not independent, all of the correlated variables get high importance scores compared to the uncorrelated ones (Strobl, AL Boulesteix, and Zeileis 2007). By using the conditional inference trees (implemented in '`cforest`' (R package 'party') by Hothorn, Hornik, and Zeileis 2006) and by applying a bootstrap resampling without replacement, the biased variable importance measure problem seemed to be improved (Strobl, AL Boulesteix, and Zeileis 2007). To definitely solve the

problem of biased variable importance measure for correlated variables, Strobl, Boulesteix, et al. (2008) proposed a new conditional variable importance measure implemented in the 'party' package. Conditional inference trees are different from normal trees (like CART) because of how the recursive binary splitting algorithm is performed. While trees result in an exhausting search through all possible two-ways splits by causing general overfitting and biased variable selection at each split, conditional inference trees separate the variable selection for splitting and the splitting procedure in this way: a $\chi^2$ test investigates the significance of the association between the outcome and one of the covariates and the covariate with strongest association is selected for splitting, then a permutation test finds the optimal binary split for that covariate. This is called *unbiased recursive partitioning* (Hothorn, Hornik, and Zeileis 2006). Conditional inference trees select the correct covariate in a split more often than the traditional exhaustive search procedure and do not need pruning as in this way they avoid overfitting.

**Lasso and Elasticnet**

The **Lasso** (Tibshirani 1996) and the **Elasticnet** (Zou and Hastie 2005) are regularized regression methods which perform shrinkage of the coefficients to enhance prediction accuracy. Unlike the regularised regression method Ridge (see Subsection 1.1.4), they perform variable selection by setting the smallest estimated coefficients exactly to 0 to reduce model complexity. Let $\boldsymbol{y}$ be the $n$-vector outcome, $\boldsymbol{X}$ the $n \times p$ design matrix of the $p$ explanatory variables and $\boldsymbol{\beta}$ the $p$-vector of least squares coefficients. The Lasso and Elasticnet estimated coefficients are given by:

$$\boldsymbol{\beta}_{lasso} = \arg\min_{\boldsymbol{\beta}} \Big\{ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \Big\} \tag{2.3}$$

$$\boldsymbol{\beta}_{enet} = \arg\min_{\boldsymbol{\beta}} \Big\{ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda [\alpha \|\boldsymbol{\beta}\|_2^2 + (1-\alpha)\|\boldsymbol{\beta}\|_1] \Big\} \tag{2.4}$$

where $\|\cdot\|_1$ and $\|\cdot\|_2$ are the $l_1$ and $l_2$ norms respectively, and $\lambda > 0$ and $0 < \alpha < 1$ are called *tuning parameters* and are selected via resampling methods (cross-validation or bootstrap) in order to minimise the MSE (or classification error) or by minimising criteria such as Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) through grid search. For example, tuning the model via bootstrapping resampling happens in the following way:

1. Choose a set of tuning parameters to evaluate
2. For each tuning parameter (or for each couple of tuning parameters as for Elasticnet) follow these steps:

   (a) draw a bootstrap sample from the model training set
   (b) fit the model on the bootstrap sample
   (c) predict the out-of-bag sample (observations not in the bootstrap sample)
   (d) repeat steps 2a-2c a large number of times (e.g. 100, Harrell, Lee, and Mark 1996)
   (e) calculate the average performance (MSE) across out-of-bag predictions

3. Determine the tuning parameter set which returns the optimal performance (minimal MSE)

4. Fit the final model to all the training data using the optimal tuning parameter set.

A cross-validation (CV) tuning procedure is similar to the above, only step 2 changes as instead of bootstrap resampling, CV is performed. In a repeated CV tuning, for each tuning parameter set CV is repeated $n$ times on $n$ different random partitions of the training data, and the final optimal tuning parameter set is given by the one returning the minimal average model MSE across the $n$ CVs (Krstajic et al. 2014).

**Lasso and Elasticnet properties** The Lasso performs more variable selection than Elasticnet returning a sparse vector of estimated coefficients. Therefore the Lasso will work more efficiently when there are large proportions of noise variables. Because the $l_1$ penalty is not strictly convex, the Lasso will only choose one variable in a group of highly correlated variables since the penalised RSS optimal solution for equal variables is not unique (Zou and Hastie 2005). In contrast, Elasticnet will select or exclude the whole group of correlated covariates as the Elasticnet penalty is convex and returns a unique solution, i.e two equal estimated coefficients for equal variables. Thus, the level of correlation of the variables and the purpose of the analysis will dictate which method is best to use between the two. Due to shrinkage, these methods are able to handle datasets with a higher number of variables compared to the number of observations ($p \geq n$). In the $p > n$ case, the Lasso selects at most $n$ variables before it saturates, because of the nature of the convex optimization problem (Zou and Hastie 2005). This limiting feature does not happen in the case of Elasticnet. The bias due to shrinkage in the parameter estimation of Lasso and Elasticnet is compensated with a reduction of variance compared to linear regression. However, there are drawbacks to these methods. One of them lies in the fact that, when the true (ordinary least squares, OLS) $\boldsymbol{\beta}$ components are larger than the chosen $\lambda$, the Lasso (and similarly the Elasticnet) estimator has a bias approximately of size $\lambda$, shifting the true coefficients larger than $\lambda$ towards 0 by a factor of size $\lambda$ (see Figure

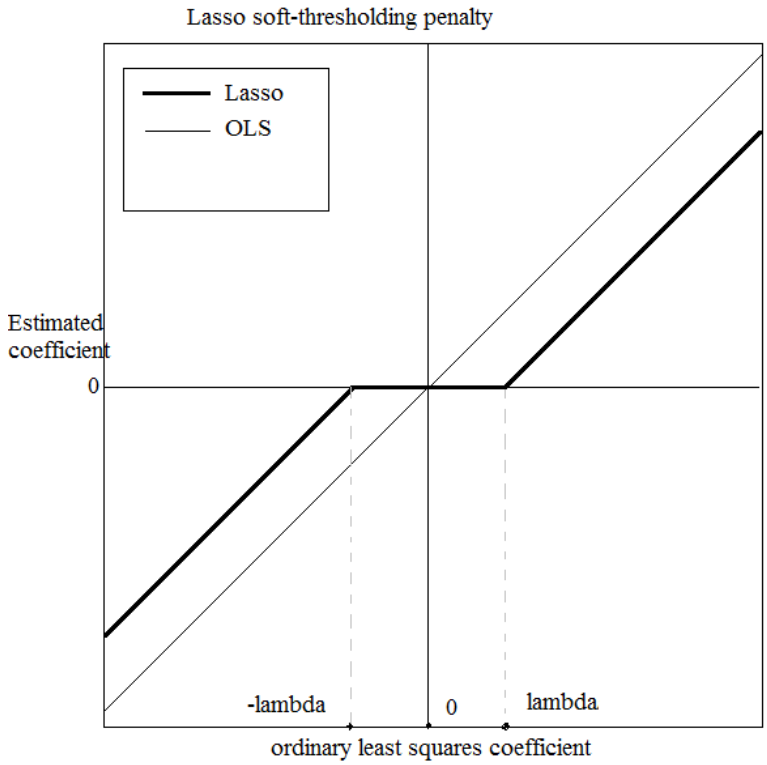



Figure 2.1: Lasso soft-thresholding penalty: the Lasso estimator is shifted towards 0 from the truth (unbiased OLS estimator) by a constant $\lambda$ when the true coefficients are larger than $\lambda$. When the true coefficients are smaller than $\lambda$, they are shrunk towards 0.

2.1, Fan and Li 2001). Therefore, when the tuning parameters are automatically selected by a data-driven rule, these regularised regression methods tend to overfit the data, by selecting more fake predictors: *model selection inconsistency* (Fan and Lv 2009 and Zou 2006). Different more complex penalties were proposed to avoid this bias, e.g. the SCAD penalty (Fan and Li 2001) and the adaptive Lasso (Zou 2006). However, a simpler solution to this overfitting problem can be choosing the Lasso (or Elasticnet) model corresponding to a stronger penalty that returns a higher MSE compared to the minimum, but with an accuracy still comparable to the optimum. This way the model is simpler and the number of false positives selected variables is lower. For example, the *one-standard-error rule* (Breiman, JH Friedman, et al. 1984, Hastie, Tibshirani, and Friedman 2008, James et al. 2013), applied in the case of cross-validation (CV) tuning, selects the most parsimonious model whose error is no more than one standard error above the error of the best model. Also Musoro et al. (2014) studied a stronger penalty other than the best for the Lasso: the 3% tolerance penalty corresponding to the model having an error within 3% of the minimum. The stronger the penalty, the more parsimonious the model.

Typically the 3% tolerance penalty results stronger than the one-standard-error penalty, unless the standard error of the minimum error is larger than the minimum error's 3% proportion.

The other drawback for regularised regression methods is the fact that they only perform *complete records analysis*, i.e. they only use the non-missing cases when missing data are present. For this reason, I considered using missing data imputation techniques (see next Subsection 2.1.2)

### 2.1.2 Missing data imputation techniques

For an overview of missing data handling methods in statistical learning see Subsection 1.2.2.

In order to take advantage of the variable selection and high dimensional data handling properties of these regularised regression methods when missing data are present, two data imputation techniques were combined with the above methods: MICE (implemented in the '`mice`' R package, Van Buuren and Oudshoorn 2000) and MissForest (implemented in the '`missForest`' R package, Stekhoven and Buhlmann 2012).

#### Multivariate imputation using chained equations (MICE)

**MICE** (Van Buuren and Oudshoorn 2000) is a classical parametric statistical method (not commonly used in statistical learning) to impute missing values that completes each variable by regressing it on all the other variables in the imputation model sequentially until all variables are complete. This process is repeated multiple times in order to incorporate missing data uncertainty. MICE can deal with mixed type of variables (continuous and categorical). MICE assumes that a full multivariate distribution exists and missing values are sampled from conditional distributions based on this full distribution. In case the rate of missing information is not high and the assumptions are met, only 5 to 10 imputations are needed to have efficient results (Rubin 1981). However, in general more imputations, depending on the percentage of missing data are recommended. MICE procedures assume that the data are at least missing at random (MAR) to deliver reliable imputation results (Van Buuren and Oudshoorn 2000). However, it is not possible to test this assumption (Carpenter and Kenward 2013).

In classical statistics, the planned analysis results obtained on the different imputed datasets are then combined using Rubin's Rules (Rubin 1981): the estimated parameters are averaged across the imputed datasets and the combined variance is given by a linear combination of the average within imputation variance and the between imputation variance. The imputation model of MICE needs to include all variables that will be investigated in the planned statistical

models, including any potential interaction that will be tested and the dependent variable of the model (KG Moons et al. 2006). Including auxiliary (i.e. additional) variables in the imputation model, which are not in the substantive model (i.e. planned analysis model), and which are predictive of missingness and predictive of the missing values of variables in the substantive model, can improve the efficiency of imputations, by reducing bias and make the MAR assumption more plausible (Hippel and Lynch 2013). When auxiliary variables are not predictive of missingness or not correlated with the variables in the substantive model, their inclusion in the imputation model can generate noise and unstable imputation estimates in small sample regression analyses (Hardt, Herke, and Leonhart 2012.

**MICE algorithm** The MICE algorithm was introduced by Van Buuren and Oudshoorn (2000) and implemented in the `mice()` R-function. The algorithm imputes missing values by Gibbs sampling: by default, each variable containing missing values is predicted from all other variables in data set. These prediction equations are used to impute plausible values for the missing data. The process iterates until convergence over the missing values is achieved.

Let us assume that $Y$ is a partially observed random sample from the $p$-variate multivariate distribution $P(\boldsymbol{Y}|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of unknown parameters. The MICE algorithm obtains the posterior distribution of $\boldsymbol{\theta}$ by sampling iteratively from conditional distributions of the form $P(\boldsymbol{Y}_1|\boldsymbol{Y}_{-1}, \theta_1), \ldots, P(\boldsymbol{Y}_1|Y_{-p}, \theta_p)$.

The first iteration of the algorithm draws missing parameters and missing values from observed marginal distributions to obtain $\theta_j*^{(1)}, j = 1, \ldots, p$ and $\boldsymbol{Y}_j*^{(1)}, j = 1, \ldots, p$. The $t-$th iteration is a Gibbs sampler that draws parameters and data from conditional distributions such as:

$$
\begin{aligned}
\theta*_1^{(t)} &\sim P(\theta_1|\boldsymbol{Y}_1^{obs}, \boldsymbol{Y}_2^{(t-1)}, \ldots, \boldsymbol{Y}_p^{(t-1)}) \\
\boldsymbol{Y}*_1^{(t)} &\sim P(\boldsymbol{Y}_1|\boldsymbol{Y}_1^{obs}, \boldsymbol{Y}_2^{(t-1)}, \ldots, \boldsymbol{Y}_p^{(t-1)}, \theta*_1^{(t)}) \\
&\vdots \\
\theta*_p^{(t)} &\sim P(\theta_p|\boldsymbol{Y}_p^{obs}, \boldsymbol{Y}_1^{(t-1)}, \ldots, \boldsymbol{Y}_{p-1}^{(t-1)}) \\
\boldsymbol{Y}*_p^{(t)} &\sim P(\boldsymbol{Y}_p|\boldsymbol{Y}_p^{obs}, \boldsymbol{Y}_1^{(t-1)}, \ldots, \boldsymbol{Y}_{p-1}^{(t-1)}, \theta*_p^{(t)})
\end{aligned}
$$

where $\boldsymbol{Y}_j^{(t)} = \left(\boldsymbol{Y}_j^{obs}, \boldsymbol{Y}_j*^{(t)}\right)$ is the $j-$th imputed variable at iteration $t$.

It is important to monitor convergence, but apparently the number of iterations can often be a small number. This process is repeated $m$ times in order to generate $m$ different imputed

datasets.

### MissForest

The non-parametric missing value imputation algorithm MissForest (Stekhoven and Buhlmann 2012) is an iterative imputation method based on Random Forests (RF) which accommodates non-linear relation structures and complex interactions, mixed type of variables (continuous and categorical), by performing well under moderate to high missingness and even (in certain cases) under the MNAR assumption (Cutler et al. 2009). MissForest intrinsically constitutes a multiple imputation scheme: for each variable in turn a RF is trained on all other variables observed values in a first step, then missing values are predicted through the built RF model, next repeatedly a RF is trained on the predicted observations, and then used to predict the missing values until the difference between the newly imputed data matrix and the previous one increases for the first time. No prior knowledge about the data is needed apart from the fact that the observations need to be pairwise independent. It outperforms MICE and other imputation methods mainly in datasets with non-linearities and with different percentages of missing values (Shah et al. 2014). Waljee et al. (2013) showed that MissForest outperformed well known methods such as k-nearest neighbours (Troyanskaya et al. 2001) and parametric MICE (Van Buuren and Oudshoorn 2000). Tang and Ishwaran (2017) revealed that MissForest was the best performing RF imputation method with performance improving with increasing correlation. One disadvantage was the longer computational time for high-dimensional data compared to the other RF missing data imputation algorithms (Tang and Ishwaran 2017). Also Shah et al. (2014) acknowledged the ability of MissForest to better manage complex data, but they showed that MissForest, as all RF imputation methods, can be biased in some situations due to the fact that it cannot impute values beyond the observed ones, not constituting a model-based prediction. However, they assessed bias and certainty of parameter estimates and not prediction accuracy.

**MissForest algorithm** Let $\boldsymbol{X}, (n \times p)$-dimensional data matrix, be our data, with $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ being the columns variables of $X$. For each variable $\boldsymbol{x}_j$ with missing values at entries $\boldsymbol{i}_{mis}^{(j)} \subseteq \{1, \ldots, n\}$, let us separate the dataset into 4 parts:

- $\boldsymbol{y}_{obs}^{(j)}$, the observed values of variable $\boldsymbol{x}_j$,
- $\boldsymbol{y}_{mis}^{(j)}$, the missing values of variable $\boldsymbol{x}_j$,
- $\boldsymbol{x}_{obs}^{(j)}$, the variables other than $\boldsymbol{x}_j$ with observations $\boldsymbol{i}_{obs}^{(j)} = \{1, \ldots, n\} \setminus \boldsymbol{i}_{mis}^{(j)}$,

- $\boldsymbol{x}_{mis}^{(j)}$, the variables other than $\boldsymbol{x}_j$ with observations $\boldsymbol{i}_{mis}$.

RF can automatically handle missing values by weighting the frequency of the observed values in a variable with the RF proximities after being trained on the initially mean imputed dataset (Breiman 2001). However, this approach requires a complete response variable for training the forest. Therefore,

1. to begin an initial guess for the missing values is made in $\boldsymbol{X}$ using mean imputation or another imputation method for mixed data (continuous and categorical variables).
2. Then, variables $\boldsymbol{x}_j, j = 1, \ldots, p$ are sorted according to the amount of missing data starting from the lowest amount.
3. For each variable $\boldsymbol{x}_j$, the missing values are imputed by first fitting an RF with response $\boldsymbol{y}_{obs}^{(j)}$ and predictors $\boldsymbol{x}_{obs}^{(j)}$.
4. Then missing values $\boldsymbol{y}_{mis}^{(j)}$ are predicted by applying the trained RF to $\boldsymbol{x}_{mis}^{(j)}$.
5. The imputation procedure is repeated until the difference between the last imputed data and the previous one increases for the first time with respect to both continuous and categorical variables (Stekhoven and Buhlmann 2012).

This simulation study combines MissForest to Lasso/Elasticnet for the first time. MissForest was already combined to Cox regression (Shah et al. 2014) and Lasso was already combined to MICE (Musoro et al. 2014) and to RF imputation by proximities (Lu and Petkova 2014. Also, other regularised regression penalties were used in combination with MICE, e.g. the Group-Lasso (Chen and Wang 2013). I particularly chose the Lasso and Elasticnet penalties for their variable selection property, their popularity and easy understanding for all audiences.

It is well known that RF imputation can produce biased estimates of missing values by predicting them in a way which produces estimates too similar to the observed data (Shah et al. 2014 and Lu and Petkova 2014). Also, prediction models applied to RF imputed data (single imputation) do not incorporate missing data uncertainty and might result too optimistic in predictions. Therefore, when prediction models are applied to RF imputed data, performance results need to be internally validated to correct for these biases (see Subsection 1.2.1).

### 2.1.3 Handling overfitting using Efron's bootstrap validation as for Harrell et al. (1996)

Internal validation procedures are used to estimate a prediction model performance on new cases belonging to the same population of the data used to train the model. As it is explained

in Subsection 1.2.1, *Bootstrap resampling* can be used as an internal validation method that returns nearly unbiased and relatively low variance estimates of future model performance. It has already been mentioned that internally validating the models through bootstrap resampling allows using all the available data to estimate the final model parameters and fewer model fits are required than CV (see Subsection 1.2.1). With bootstrap validation, one repeatedly fits the model in a bootstrap sample and evaluates the performance of the model on the original sample. The estimate of the likely performance of the final model on new data is estimated by the average of all the bootstrap sample model estimates of accuracy computed on the original sample that is slightly biased downward. Efron (1979, pages 247-252) improved the accuracy of the bootstrap estimate of model performance by estimating the optimism in the final model fit and by subtracting it from the apparent performance derived from the original sample to obtain a bias-corrected (overfitting-corrected) estimate of performance. The following are the steps needed to run an *Efron's optimism bootstrap validation* for a given model as for Harrell, Lee and Mark (1996):

1. Develop the model $M$ using all $n$ observations and measure the apparent performance measure of interest $P_{apparent}$ computed on the same $n$ observations used to develop $M$.
2. Draw a bootstrap sample from the original sample.
3. Train the model on the bootstrap sample naming it model $M_{boot}$.
4. Compute the apparent (training) performance for model $M_{boot}$ on the same bootstrap sample denoting it $P_{boot}$.
5. Compute the performance of model $M_{boot}$ on the original data and denote it $P_{test}$
6. The optimism in the fit from the bootstrap sample is $P_{boot} - P_{test}$
7. Repeat steps 2 to 6 at least 100 times
8. Average all the optimism estimates from the bootstrap samples to obtain the internal optimism $O_{internal}$.
9. The bootstrap corrected performance of the original model $M$ is $P_{corrected} = P_{apparent} - O_{internal}$, nearly unbiased estimate of the expected value of the external performance of the process that generated $M$.

Therefore, $P_{corrected}$ is an honest estimate of internal validity that penalizes for overfitting.

Bootstrapping can be used to estimate optimism in any performance measure. For example, bias-corrected estimates of *calibration slope* and *calibration-in-the-large* measures can also be obtained via Efron's optimism bootstrap (see Subsection 1.2.1).

Bootstrapping is here preferred to cross-validation as an internal validation method because of the fewer resampling steps required (and then cheaper computational cost) in order to compute nearly unbiased estimates of optimism (Harrell, Lee, and Mark 1996 and Steyerberg 2009).

**Some details on Efron's optimism bootstrap as for Harrell, Lee and Mark (1996)**

In this paragraph I will restate Efron's optimism bootstrap procedure (Harrell, Lee, and Mark 1996) in precise language using the theory of probability (Efron 1979, Hastie, Tibshirani, and Friedman 2008, Borra and Di Ciaccio 2010). As for the notation used in Subsection 1.2.1, let us call $\mathcal{T}_i = (\boldsymbol{x}_i, y_i), i = 1, \ldots, n$ the observations of the individual $i$, where $\boldsymbol{x}_i$ is the vector of inputs and $y_i$ is the outcome. Efron's bootstrap procedure considers $\mathcal{T} = \{(\boldsymbol{x}_1, y_2), \ldots, (\boldsymbol{x}_n, y_n)\}$ as an i.i.d. sample from the multidimensional distribution $F$. If the model $\hat{f}_{\mathcal{T}}$ is estimated from our data $\mathcal{T}$, then $\hat{\boldsymbol{y}} = \hat{f}_{\mathcal{T}}(\boldsymbol{x}_0)$ is the predicted value of $y$ at $\boldsymbol{x} = \boldsymbol{x}_0$. Let us consider $\hat{f}_{\mathcal{T}}(\boldsymbol{x}_0)$ as a plug-in statistic $\hat{f}_{\mathcal{T}}(\boldsymbol{x}_0) =: \hat{f}(\boldsymbol{x}_0, \hat{F})$ for some function $\hat{f}$, where $\hat{F}$ is the empirical distribution function of the data. We are here relying on the assumption that the empirical distribution of the sample (the set of all the possible $n^n$ bootstrap samples from the sample under interest) converges to the true distribution $F$ as the sample size becomes large. The bootstrap using all of the $n^n$ samples is called the ideal bootstrap. However, the bootstrap generally only draws a number of the $n^n$ samples randomly with replacement, and approximates the ideal bootstrap as the the number of boostrap samples becomes large. Let us call $L[y, \hat{f}(\boldsymbol{x})]$ a measure of error between the response $y$ and the prediction $\hat{f}(\boldsymbol{x})$. Thus, the *prediction error* (also called *true error*, *generalization error* or *extra-sample error*) for $\hat{f}_{\mathcal{T}}(\boldsymbol{x}_0)$ is defined:

$$\text{Err}_{\mathcal{T},F} := E_{0,F}\left[L[y_0, \hat{f}_{\mathcal{T}}(\boldsymbol{x}_0)|\mathcal{T}]\right], \tag{2.5}$$

where the notation $E_{0,F}$ means the expectation over a new observation $(\boldsymbol{x}_0, y_0)$ from the population $F$ and $\mathcal{T}$ is fixed. $\text{Err}_{\mathcal{T},F}$ refers to the error for this specific training set $\mathcal{T}$, it is a conditional error. The *apparent error* (also called *training error*) is defined as follows:

$$\text{err}_{\mathcal{T},\hat{F}} := E_{\hat{F}}\left[L[y, \hat{f}_{\mathcal{T}}(\boldsymbol{x})|\mathcal{T}]\right] = \frac{1}{n}\sum_{i=1}^{n} L[y_i, \hat{f}_{\mathcal{T}}(\boldsymbol{x}_i)]. \tag{2.6}$$

It is called the apparent error because $E_{\hat{F}}$ averages on the losses on the training data $\mathcal{T}$ and the same data is used to fit the method and assess its error.

The bootstrap procedure can estimate the true error $\text{Err}_{\mathcal{T},F}$ (2.5) by applying the plug-in principle (for which the true distribution $F$ is approximated by the empirical distribution $\hat{F}$, consequence of Slutsky's theorem) as follows:

$$\widehat{\text{Err}}_{\mathcal{T}^*,\hat{F}} = \frac{1}{n}\sum_{i=1}^{n} L[y_i, \hat{f}_{\mathcal{T}^*}(\boldsymbol{x}_i)], \tag{2.7}$$

where $\mathcal{T}^* = \{(\boldsymbol{x}_1^*, y_1^*), \ldots, (\boldsymbol{x}_n^*, y_n^*)\}$ is a bootstrap sample, $\hat{f}_{\mathcal{T}^*}(\boldsymbol{x}_i)$ is the predicted value at $\boldsymbol{x} = \boldsymbol{x}_i$, according to the model $\hat{f}_{\mathcal{T}^*}$ trained on the bootstrap sample $\mathcal{T}^*$. However, this estimate of the true error is given for a fixed single bootstrap sample and this makes it to have high variance across bootstrap samples. Let us then define the *expected extra-sample error* (or *average prediction error* or *expected prediction error*):

$$\text{Err}_F := E_{\mathcal{T},F}\left[\text{Err}_{\mathcal{T},F}\right] = E_{\mathcal{T},F}\left[E_{0,F}\left[L[y_0, \hat{f}_{\mathcal{T}}(\boldsymbol{x})|\mathcal{T}]\right]\right] \tag{2.8}$$

that averages over new observations and training sets $\mathcal{T}$ drawn randomly. The ideal bootstrap plug-in estimate of this quantity is:

$$\widehat{\text{Err}}_{\hat{F}} = E_{\mathcal{T}^*,\hat{F}}\left[\text{Err}_{\mathcal{T}^*,\hat{F}}\right] = E_{\mathcal{T}^*,\hat{F}}\left[\sum_{i=1}^{n} L[y_i, \hat{f}_{\mathcal{T}^*}(\boldsymbol{x}_i)]/n\right], \tag{2.9}$$

since it averages over an infinite number of bootstrap samples (or over the $n^n$ possible combinations of the empirical distribution $\hat{F}$). Therefore, an approximation corresponding to a finite number of bootstrap samples needs to be applied:

$$\widehat{\text{Err}}^*_{\hat{F}} = \frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{n} L[y_i, \hat{f}_{\mathcal{T}_b^*}(\boldsymbol{x}_i)]/n. \tag{2.10}$$

The bootstrap approach estimates the bias in the apparent error $\text{err}_{\mathcal{T},\hat{F}}$ (2.6) as an estimator of the true error $\text{Err}_{\mathcal{T},F}$ and then corrects $\text{err}_{\mathcal{T},\hat{F}}$ (2.6) by subtracting its estimated bias. This bias is called *optimism*:

$$\text{op} := \text{Err}_{\mathcal{T},F} - \text{err}_{\mathcal{T},\hat{F}} \tag{2.11}$$

It is important to note that the optimism of the apparent error can also be defined by focusing on a more restrictive estimate of the prediction error: the *in-sample error*, i.e. the expectation of the losses according to a fixed function $\hat{f}_{\mathcal{T}}$ over the new possible outcomes while the covariates

are considered fixed at their observed samples (Hastie, Tibshirani, and Friedman 2008, Borra and Di Ciaccio 2010). However, the in-sample error is not discussed further in the thesis as the extra-sample error is used instead.

Let us define now the *average optimism* (*expected optimism*):

$$\boldsymbol{\omega}_F := E_{\mathcal{T},F}[\text{op}] = E_{\mathcal{T},F}\left[\text{Err}_{\mathcal{T},F} - \text{err}_{\mathcal{T},\hat{F}}\right], \tag{2.12}$$

i.e. the average difference between the true prediction error and the apparent error over data sets $\mathcal{T}$ with pairs of observations $\mathcal{T}_i = (\boldsymbol{x}_i, y_i) \sim F$. With this definition, it will usually turn out that the expected optimism $\boldsymbol{\omega}_F$ (2.12) is positive as the apparent error rate $\text{err}_{\mathcal{T},\hat{F}}$ is downward biased in estimating the prediction error. Harrell, Lee, and Mark 1996 and Steyerberg 2009 define the optimism in the opposite way: op:$= \text{err}_{\mathcal{T},\hat{F}} - \text{Err}_{\mathcal{T},F}$, thus it will be negative most of the times (see section 2.1.3). However, these definitions are equivalent except for the sign.

The plug-in bootstrap estimate of the expected optimism $\omega_F$ is:

$$\hat{\boldsymbol{\omega}}_{\hat{\boldsymbol{F}}} = E_{\mathcal{T}^*,\hat{F}}\left[\widehat{\text{Err}}_{\mathcal{T}^*,\hat{F}} - \widehat{\text{err}}_{\mathcal{T}^*,\hat{F}^*}\right], \tag{2.13}$$

where the $\hat{F}^*$ is the empirical distribution function of the bootstrap sample $\mathcal{T}^*$. This ideal bootstrap quantity is approximated by

$$\hat{\boldsymbol{\omega}}^*_{\hat{F}} = \frac{1}{nB}\sum_{b=1}^{B}\sum_{i=1}^{n}\{L[y_i, \hat{f}_{\mathcal{T}_b^*}(\boldsymbol{x}_i^*)] - L[y_{ib}^*, \hat{f}_{\mathcal{T}_b^*}(\boldsymbol{x}_i^*)]\}, \tag{2.14}$$

where $\hat{f}_{\mathcal{T}_b^*}(\boldsymbol{x}_i^*)$ is the predicted value at $\boldsymbol{x}_i^*$ of the model trained on the $bth$ bootstrap sample, $b = 1, \ldots, B$, and $y_{ib}^*$ is the outcome of the $i$-th observation for the $b$-th bootstrap sample, $b = 1, \ldots, B$. Therefore, the final estimate of prediction error is then given by:

$$\widehat{\text{Err}}_{\mathcal{T},F} = \text{err}_{\mathcal{T},\hat{F}} + \boldsymbol{\omega}_{\hat{F}}, \tag{2.15}$$

i.e. the apparent error plus the bias of the apparent error (the expected optimism). The bootstrap approximation of the estimated prediction error (2.15) is:

$$\widehat{\text{Err}}^*_{\mathcal{T},\hat{F}} = \text{err}_{\mathcal{T},\hat{F}} + \hat{\boldsymbol{\omega}}^*_{\hat{F}}. \tag{2.16}$$

This bootstrap estimate of the prediction error (2.15) for a large number of bootstrap samples is nearly unbiased (Efron and RJ Tibshirani 1994).

These details on Efron's optimism bootstrap algorithm meant to formalise the concept of optimism correction of the estimated extra-sample error. Optimism correction can also be applied to all performance measures, e.g. calibration and discrimination measures.

K-fold CV can perform optimism correction yielding similar nearly unbiased estimates of optimism, but with larger variance than bootstrap. In order to achieve stable estimates, a larger number of resampling steps are required through repeated CV, which results in longer computational time.

I will apply optimism-correction of discrimination and calibration performance measures to the combined methods assessed in my simulation study.

### 2.1.4 Monte Carlo simulation study

A Monte Carlo (MC) simulation study is a numerical technique in statistics for conducting experiments on the computer involving random sampling from probability distributions (Burton et al. 2006). In the traditional frequentist inferential approach of statistical modelling, MC simulations are typically used to assess the properties of estimators, hypothesis tests and confidence intervals, so that the methods can be used with confidence. For example, simulations are used to assess the bias of an estimator in finite samples, its consistency under departures from assumptions, its sampling variance and how it compares to competing estimators in terms of bias and precision, or to assess whether a constructed confidence interval for a parameter achieves the established nominal level of coverage (Boos and Stefanski 2013). Exact analytical derivations of these properties are rarely possible (Harrison 2010).

MC simulations allow to approximate the sampling distribution of an estimator under particular conditions (e.g. finite sample size, true distribution of the data) in order to address the issues above, which are analytically intractable or for which the experimentation is costly or too time-consuming or not feasible.

**MC simulations in classical statistics** A classical statistics MC simulation study, that approximates the true value $\theta_0$ of an estimator $\hat{\theta}$, involves the following steps:

1. Generate $S$ independent datasets under the assumptions of interest,
2. Compute the numerical value of the estimator $\hat{\theta}$ from each dataset: $\hat{\theta}_1, \ldots, \hat{\theta}_S$,
3. Compute the mean $\bar{\theta}$ of $\hat{\theta}_1, \ldots, \hat{\theta}_S$ that will be the MC approximation of the true value of $\hat{\theta}$

under the assumptions of interest (if $S$ is large enough):

$$\widehat{mean} = S^{-1}\sum_{s=1}^{S}\hat{\theta}_s = \bar{\theta}, \qquad \widehat{SD} = \sqrt{(S-1)^{-1}\sum_{s=1}^{S}(\hat{\theta}_s - \bar{\theta})^2}, \tag{2.17}$$

where $\widehat{SD}$ is also called the empirical standard error of the simulation estimate.

If using the mean and SD of the estimates over all simulations is not considered appropriate, then non-parametric summary measures using quantiles of the distribution could be obtained.

When $K$ different estimators $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(K)}$ for the true parameter $\theta_0$ of a distribution based on i.i.d. draws $Y_1, \ldots, Y_n$ for the random variable $\boldsymbol{Y}$ need to be compared, the following estimated performance measures for each estimator will assess them:

- *bias*: $\widehat{bias}^{(k)} = \bar{\theta}^{(k)} - \theta_0$,
- *accuracy*: $\widehat{MSE}^{(k)} = \widehat{SD^{(k)}}^2 + (\widehat{bias}^{(k)})^2$,
- *coverage* for estimator $\hat{\theta}^{(k)}$, i.e. the proportion of times the $100(1-\alpha)\%$ confidence interval $\hat{\theta}_s^{(k)} \pm Z_{1-\alpha/2}\widehat{SE}(\hat{\theta}_s^{(k)}), s = 1, \ldots, S$ include $\theta_0$, $\alpha$ being the level of significance chosen, $Z_{1-\alpha/2}$ being the $1-(\alpha/2)$ quantile of the standard normal distribution and $\widehat{SE}(\hat{\theta}_s^{(k)}), s = 1, \ldots, S$ is the standard error of the estimate of interest within each simulation.

Simulations are *fully independent* when they involve generating a completely different set of independent draws $Y_1, \ldots, Y_n$ ($n$-sized sample) for the random variable $\boldsymbol{Y}$ for each estimator $\hat{\theta}^{(k)}$ and scenario considered (i.e varying sample size, correlation between variables, etc.). Simulations are *moderately independent*, when the same set of simulated independent draws are used to compare different estimators for the same scenario (Burton et al. 2006).

The suitable number of simulations $S$ depends on the true value $\theta_0$ of $\theta$ the parameter of interest, on the variance of $\theta$ (that may be known from asymptotic theory or preliminary runs), and on the required accuracy $\delta$, i.e. the permissible difference from the true value (Burton et al. 2006):

$$S = \frac{Z_{1-\alpha/2}Var(\theta)}{\delta}. \tag{2.18}$$

**MC simulations in statistical learning** In statistical learning the above method for assessing simulations cannot be used in most applications, because it is impossible to obtain a sufficiently precise estimate of the bias for most methods (e.g. Lasso, Elasticnet and Random Forests, see Subsection 2.1.1). Reliable estimates of the bias are only available if reliable unbiased estimates are available, which is typically not the case for some penalized estimates (Tibshirani

1996). The main target parameters of the present simulation study are (1) the extra-sample prediction error (estimated by the MSE of the predictions, see Subsections 1.2.1 and 2.1.4, usual target in statistical learning prediction modelling) and (2) the percentage of selected true predictors. The statistical learning models I aim to analyse (Lasso, Elasticnet and Random Forest) do not have closed theoretical formulas for the prediction error as, for example, does linear regression.

Therefore, the present simulation study comparing different methods was not assessed in the classical way. The estimators of the main target parameter, the MSE of the predictions, do not have a known true value and so bias, accuracy (i.e. the MSE of the MSE) and coverage of each estimator were not assessed. Instead, by assuming a true linear regression model for the dependent variable, through simulations, I saw how much the prediction error of a regularised model (or Random Forests) departed from the theoretical MSE given by the true linear regression model. To have an estimate of uncertainty, I looked at the interval given by the 2.5th and 97.5th percentiles of the parameter simulation distribution (or its empirical standard error across simulations, i.e. the standard deviation of the estimates, for the percentages of selected true predictors). Confidence intervals could not be computed, because the SE of the estimate of interest within each simulation step is not very meaningful as penalized estimation methods produce biased estimates (Fan and Lv 2009).

These will be the generic steps performed in the simulations:

1. Generate $S$ independent datasets with $K$ standardised covariates each, having a given correlation structure, and generate a dependent variable from the $K$ covariates according to a multiple linear regression model,

2. Run the statistical learning models of interest on each dataset

3. Compute the numerical value of the target parameters from each model and each dataset.

4. Compute the mean of the measured target parameters across simulations and record the 2.5th and 97.5th percentiles of the simulation distribution (or its empirical standard error) to have an estimate of uncertainty.

5. Compare the means of target parameters between models and with true multiple linear regression model parameter.

### 2.1.5 Hypotheses

In this simulation study, I aimed to compare validated prediction accuracy and variable selection performance of Lasso, Elasticnet, Random Forests and Random Forests with conditional inference trees (which I will call in short Conditional Random Forest) when missing data were imputed through MICE or MissForest. I expected the following general results:

1. Lasso would need a stronger penalty to correct for model selection inconsistency with increasing ratio between number of covariates and sample size (Fan and Lv 2009)

2. Elasticnet would outperform Lasso when predictors are highly correlated (Zou and Hastie 2005)

3. Random Forests would perform best in prediction accuracy but would often consider more important the continuous variables than the categorical ones (Strobl, Boulesteix, et al. 2008);

4. Random Forest and MissForest performances would improve with increasing correlation between the variables (Tang and Ishwaran 2017);

5. MICE combined with Lasso as for Musoro et al. 2014 would return good variable selection for a penalty $\lambda$ corresponding to a model having the MSE within 3% of the minimum and not for the best $\lambda$ returning the model with the minimum MSE;

6. MissForest combined with Lasso would prevent the noise terms with large missingness from being selected (this would not happen when true predictors were strongly correlated between them, Lu and Petkova 2014);

7. MissForest combined with Random Forests would tend to give more often importance to the noise variables with large probability of missingness (Cutler et al. 2009);

8. MissForest would outperform MICE (when combined with Lasso) with increasing ratio noise to true predictors variables in the imputation model (Hardt, Herke, and Leonhart 2012)

9. Prediction accuracy would decrease and optimism increase when missing data were present also in the dependent variable (Chen and Wang 2013);

10. The performances of the methods would be equivalent with MAR or MCAR missing data (Chen and Wang 2013).

## 2.2 Methods

The six combined methods analysed in the simulation study were:

- **MICE-Lasso** (Musoro et al. 2014): MICE (Van Buuren and Oudshoorn 2000) was applied to the original incomplete dataset (included the outcome) and 10 imputed datasets were returned. Then the Lasso (Tibshirani 1996) was run on each of the 10 imputed datasets with *bootstrap resampling tuning* for $\lambda$: given a grid of 40 penalty values $\lambda$, for each $\lambda$ a Lasso model was constructed on a bootstrap sample (drawn randomly with replacement and of the same size of that imputed data set) and a bootstrap corrected mean squared error (MSE) was computed by comparing the observed and predicted values in the imputed data set. This was repeated 100 times for each penalty value and the average MSE was computed. The best penalty was the one returning the lowest average MSE. Four models per imputed data set were retained: the one corresponding to the optimal penalty (denoted as *best*) and three *tolerance* models. One tolerance model corresponded to a penalty that had an MSE within 1 standard error of the minimum MSE (denoted as *1SE*, applying the *'one-standard-error rule'* as for Breiman, JH Friedman, et al. 1984, Hastie, Tibshirani, and Friedman 2008, James et al. 2013) and the other 2 models had MSEs within 3% and 15% of the minimum respectively, yielding more parsimony (Musoro et al. 2014 only analysed the 'best' and the 3% tolerance models). These four penalties models were referred to as:

  - Best model
  - 1 SE tolerance model
  - 3% tolerance model
  - 15% tolerance model

  The final best and tolerance models had regression coefficients that were the averages over the ten imputed data sets (see flowchart 1.3);

- **MICE-Elasticnet**: it used the same procedure as MICE-Lasso with the Lasso replaced by the Elasticnet penalty (Zou and Hastie 2005) and the grid of penalty values being of length 20 for the tuning parameter $\lambda$ and 9 for the tuning parameter $\alpha$ (0.1 to 0.9) for a total of 180 combined penalty values (Lasso was not allowed to be selected). Because of the larger number of tuning parameter sets, this simulation was computationally more expensive;

- **MissForest-Lasso and MissForest-Elasticnet**: they consisted of applying the Lasso/Elasticnet as above on the original data imputed once by MissForest (Stekhoven and Buhlmann 2012).

- **MissForest-Random Forests(RF) and Conditional RF**: they consisted of applying RF or Conditional RF on the original data imputed once by MissForest.

The combined methods to compare were run on the simulated datasets as for the scenario under interest. The performance of the methods was evaluated in terms of prediction accuracy, through discrimination and calibration (see section 1.2.1), and variable selection. For each method, **discrimination** was assessed in the following way:

1. the model (developed on the whole dataset) was evaluated on the original sample to obtain an estimate of apparent discrimination, $\text{MSE}_{apparent} = \sum_{i=1}^{n}(\hat{y} - y)^2/n$, where $n$ is the sample size, $\boldsymbol{y}$ is the simulated outcome and $\hat{\boldsymbol{y}}$ the predicted outcome;

2. then the expected optimism for the MSE, referred to as *internal optimism* ($\text{Optimism}_{internal}$, see Table 2.1), was estimated for each method-scenario combination through bootstrap internal validation as described by Harrell 2001 (see section 2.1.3);

3. an estimate of internally validated discrimination for the model was given by the average across simulations of the optimism-corrected *mean squared error* ($\text{MSE}_{corrected}$) and corresponding pseudo-$R^2_{corrected}$ = 1-$\text{MSE}_{corrected}$/Var(outcome) (Breiman 2001). In order to compute the pseudo-$R^2_{corrected}$, the empirical mean outcome variance across the simulated datasets was used.

4. finally, the models were evaluated on a new independent complete simulated dataset to obtain an estimate of external discriminative performance, $\text{MSE}_{external}$.

5. The observed optimism of the model performance on external data was the difference between $\text{MSE}_{apparent}$ and $\text{MSE}_{external}$($\text{Optimism}_{external}$). $\text{Optimism}_{internal}$ and $\text{Optimism}_{external}$ are expected to be close if the resampling procedure gives unbiased estimates of optimism.

**Calibration** was measured through the average calibration slope of the linear predictor (LP) of the model built on the original data, $\beta_{LP}$. Because the methods under interest do not return unbiased estimates of the coefficients, $\beta_{LP}$ cannot be exactly 1, but it constitutes an apparent measure of calibration (Musoro et al. 2014).

**Variable selection** for the Lasso and Elasticnet was measured through:

- Sensitivity of selection ($SEN$): mean percentages of the selected *true predictors* (TPs) among the TPs
- False-positive rate of selection ($FPR$): mean percentages of the selected *noise variables* (also called fake or false predictors, FPs) among the FPs
- Positive predictive value of selection ($PPV$): mean percentages of the selected TPs among the selected variables:

I also calculated the percentage of true models selected across simulations by the Lasso and the Elasticnet and the individual inclusion frequency of the variables in the models. The variable individual inclusion frequency for Random Forest was as follows: a variable was considered included in the model when its importance was among the top 10 variable importances. I finally looked at the percentage of times the TPs were among the $n$ top coefficient variables (with $n$ being the number of TPs) as ranked by the model (for all analysed models included RF and conditional RF).

The criterion for good prediction accuracy was an $\text{MSE}_{corrected}$ close to the theoretical MSE of the true linear regression generative model for the dependent variable, which I considered a minimal optimal bound, with corresponding internal and external estimates of optimism being close to each other. Values for the $\text{MSE}_{corrected}$ departing from the theoretical MSE by more than 30% of the theoretical MSE were subjectively considered poor.

Based on discussion with clinicians, I defined the following subjective criteria for good variable selection for regularised regression methods:

- SEN $\leq$ 70% poor
- SEN $>$ 70% and 60% $<$ PPV $<$ 70% acceptable
- SEN $>$ 70% and 70% $\leq$ PPV $<$ 90% good
- SEN $>$ 70% and PPV $\geq$ 90% very good

However, subjective classification of poor, acceptable or good results might differ in different settings and, in the early stage of developing a model, lower performance can still be acceptable. RF or Conditional RF had good variable selection performance when the individual TP frequency of being among the first $n$ important variables (with $n$ being the number of TP) ranged from 65% to 100% and at the same time the FP individual frequency ranged between 0% to 30%.

Table 2.1: Definition of performance measures

| Measure | Definition |
|---|---|
| $MSE_{apparent}$ | Performance on the original dataset of the *model* trained on the same original dataset |
| $MSE_{external}$ | Performance of the model on an independent new complete dataset |
| $Optimism_{internal}$ | mean($MSE_{boot}$-$MSE_{test}$). Average optimism of the bootstrap model performances on the original data, i.e. average difference between the bootstrap performances (on bootstrap data) and the test performances (on original data) |
| $Optimism_{external}$ | $MSE_{apparent} - MSE_{external}$. Optimism of the model performance on external data |
| $\beta_{LP}$ | Calibration slope, i.e. slope of the linear predictor (LP) of the model estimated from regressing the observed outcome on the LP from the original data |
| $\beta_{LP^*}$ | Optimism-corrected calibration slope of the linear predictor through bootstrap validation. |

As a measure of variability for the estimates, a 2.5th to 97.5th percentile range across simulation runs was returned for each estimate, instead of returning a confidence interval based on biased estimates (see Subsection 2.1.4). The variability for SEN, FPR and PPV was also measured by sample standard deviation across simulations.

As I was aiming to accurately predict the outcome rather than to estimate the coefficients in an unbiased way, I did not measure the coefficient estimates' bias.

The method with the best trade-off between good prediction accuracy and good variable selection performances was chosen.

Because of the computational intensity of the analysed combined methods and the subsequent large time required to run them, I only simulated 300 datasets per scenario.

The simulations were moderately independent (Burton et al. 2006), i.e. I used the same set of simulated independent datasets to compare the different combined methods for the same scenario in order to detect any differences between methods.

I simulated data sets with i) 20 covariates (250 and 1000 observations) and ii) 100 covariates (500 observations). For each combination, I simulated MCAR and MAR missingness pattern with different percentages of missing data and different correlation matrices. To simulate an RCT where treatment response was different between treatment and control arm, I simulated interactions between a binary variable (treatment arm) and other variables (baseline predictors). The imputation models of MICE and MissForest included all the variables in the substantive model (planned analysis), comprising the outcome, when there was no moderation assumption. On the contrary, when some interaction terms were true predictors of the outcome,

imputation and substantive models were different: the interaction terms were not included in the imputation model (apart from MICE-Lasso, see below). MissForest imputation models did not contain interactions because RF imputation techniques already account for existing interactions in the data (Cutler et al. 2009 and Stekhoven and Buhlmann 2012). Missingness of the predictor variables was simulated not to depend on the interaction terms.

In the following Subsections 2.2.1 and 2.2.2, I will present the studied scenarios in details and justify the choices.

### 2.2.1 20-Covariate Dataset

I ran the 20 covariate simulation study as for Musoro et al. 2014 (but with additional settings) in order to replicate their results and compare their method to the other three proposed methods in this study. Firstly, 20 continuous covariates $\boldsymbol{X}_j, j = 1, \ldots, 20$ with the number of observations $n$ being both 250 and 1000, were drawn from a multivariate standard normal distribution $N(\mathbf{0}, \boldsymbol{P})$, where the correlation matrix $\boldsymbol{P}$ was sparse with the following non-zero correlations:

$$\begin{aligned} \rho_{1,5} &= 0.72 \\ \rho_{1,6} &= -0.52 \\ \rho_{2,8} &= 0.74 \\ \rho_{4,12} &= -0.82 \\ \rho_{6,16} &= -0.34 \\ \rho_{10,20} &= -0.38 \\ \rho_{11,19} &= 0.37 \\ \rho_{19,20} &= -0.65 \end{aligned} \tag{2.19}$$

Then variables $\boldsymbol{X}_1$ to $\boldsymbol{X}_{10}$ were dichotomized to binary variables at the following percentiles:

- $\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_6, \boldsymbol{X}_7$ at their 50th percentile
- $\boldsymbol{X}_3, \boldsymbol{X}_4, \boldsymbol{X}_8, \boldsymbol{X}_9$ at their 30th percentile
- $\boldsymbol{X}_5, \boldsymbol{X}_{10}$ at their 20th percentile

Two settings were considered for the generation of the outcome:

1. *No assumption of moderation*: the continuous outcome $\boldsymbol{y}$ was generated by 5 continuous

and 5 binary variables from the following linear regression model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.20}$$

where $\boldsymbol{X} = (\mathbf{1}, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_{20})$ was the design matrix of explanatory variables which columns were the constant vector $\underline{1}$ for the intercept and the 20 covariates $\boldsymbol{X}_j, j = 1, \ldots, 20$, $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\boldsymbol{I})$, $\sigma = 1.74$, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{20})$ was the vector of the regression coefficients:

$$\begin{aligned}
\beta_0 &= 1.140 \\
\beta_i &= 0, i = 1, \ldots, 5 \\
\beta_6 &= -0.839 \\
\beta_7 &= 1.131 \\
\beta_8 &= -1.540 \\
\beta_9 &= 1.426 \\
\beta_{10} &= 0.854 \\
\beta_k &= 0, k = 11, \ldots, 15 \\
\beta_{16} &= 0.457 \\
\beta_{17} &= -0.494 \\
\beta_{18} &= -0.738 \\
\beta_{19} &= 1.589 \\
\beta_{20} &= 0.845
\end{aligned}$$

2. *Interactions between predictors and a binary variable, assumption of moderation*: to assess the ability of the model to recognize moderator variables, interaction terms between the first binary variable $\boldsymbol{X}_1$ (now turned into a weak predictor, $\beta_1 = 0.1$) and some of the covariates were added to the predictors:

   - $\boldsymbol{X}_1 \cdot \boldsymbol{X}_7$, with $\beta_{1,7} = 0.6$,
   - $\boldsymbol{X}_1 \cdot \boldsymbol{X}_{13}$, with $\beta_{1,13} = 0.9$,
   - $\boldsymbol{X}_1 \cdot \boldsymbol{X}_{16}$, with $\beta_{1,16} = 1.0$,

- $\boldsymbol{X}_1 \cdot \boldsymbol{X}_{20}$, with $\beta_{1,20} = 1.2$.

SEN, FPR and PPV for the selection of the true interaction terms only were also recorded to assess the model identification of potential moderators.

Musoro et al. 2014 only considered the setting with no assumption of moderation.

Finally, eight out of the 20 covariates were set to have missing values both MCAR and MAR. In the case of MCAR data, observations to be missing were drawn from a binomial distribution with probabilities 0.2 for $\boldsymbol{X}_2, \boldsymbol{X}_7, \boldsymbol{X}_{12}, \boldsymbol{X}_{17}$ and 0.5 for $\boldsymbol{X}_3, \boldsymbol{X}_8, \boldsymbol{X}_{13}, \boldsymbol{X}_{18}$. For MAR data, the probability of missingness depended on the non-missing covariates (apart from variable $\boldsymbol{X}_5$) according to two logistic models giving average percentages of missingness of approximately 20% and 50% for the same variables:

- the probability of 20% was given by $1+\exp(-0.83\underline{\mathbf{1}}-0.79\boldsymbol{X}_1-0.35\boldsymbol{X}_4-0.89\boldsymbol{X}_6-1.51\boldsymbol{X}_9-2.01\boldsymbol{X}_{10}-2.75\boldsymbol{X}_{11}-2.06\boldsymbol{X}_{14}+1.39\boldsymbol{X}_{15}-3.44\boldsymbol{X}_{16}-0.89\boldsymbol{X}_{19}+0.67\boldsymbol{X}_{20})$,
- the probability of 50% was given by $1+\exp(-0.56\underline{\mathbf{1}}-0.69\boldsymbol{X}_1+0.36\boldsymbol{X}_4-0.25\boldsymbol{X}_6-0.16\boldsymbol{X}_9+0.86\boldsymbol{X}_{10}+0.26\boldsymbol{X}_{11}-2.27\boldsymbol{X}_{14}-1.19\boldsymbol{X}_{15}-0.13\boldsymbol{X}_{16}-1.08\boldsymbol{X}_{19}-2.12\boldsymbol{X}_{20})$.

Table 2.2 shows the simulation scenarios considered in the study for the 20 covariate datasets.

Table 2.2: Simulation study scenarios

| Scenario | Description |
|---|---|
| S1 | No missing data, No assumption of moderation |
| S2 | Missing data (MCAR and MAR), No assumption of moderation, complete outcome |
| S3 | No missing data, Assumption of moderation |
| S4 | Missing data (MCAR and MAR), Assumption of moderation, complete outcome |
| S5 | Missing data in outcome (20% missingness MCAR and MAR) |
| S6 | Interactions in the imputation model, complete outcome (only for MICE-Lasso) |

In the 20 covariate dataset study, because of the correlational structure, I did not perform simulations with the computational expensive MissForest-Elasticnet model. As no sets of correlated variables were simulated (there were not highly correlated true predictors, there was only high correlation between fake and true predictors, see equation (2.19)), Elasticnet would perform less well than Lasso (Zou and Hastie 2005).

For the 20 covariate dataset scenario, Random Forests (RF) was run only without conditional trees since the correlation matrix of the variables was sparse and the there were only continuous and binary variables. The tuning parameter given by the number of variables chosen

randomly at each split to build the trees (see Subsection 2.1.1) was chosen through bootstrap tuning (100 bootstrap datasets).

**Minor analyses** Four secondary analyses were conducted:

1. Scenarios S1 and S2 (see Table 2.2) were also analysed for MICE-Lasso with a 20 times repeated 10-cross-validation tuning procedure (the number of repetitions was suggested by Steyerberg 2009 and Harrell, Lee, and Mark 1996 to achieve stable estimates of MSE) to check whether there was an advantage in terms of prediction accuracy compared to using bootstrap tuning.

2. MICE-Lasso was also studied in the scenarios S1 and S2 with a bootstrap validation where each model, trained on a bootstrap sample, was tested on the OOB observations (observations not drawn in the bootstrap sample) and not on the whole original sample, in order to compare the average test MSEs with the optimism-corrected MSE estimates from the bootstrap validation procedure as for Harrell, Lee and Mark (1996). By testing on the OOB samples, some of the optimism in the MSE estimate should be avoided and the optimism correction would not be needed (James et al. 2013).

3. I also tried a 'majority method' for MICE-Lasso in the scenario S2 (Heymans et al. 2007): instead of getting the final model coefficients by averaging the selected variables coefficients in each of the 10 imputed datasets, I retained a variable in the final model with the following 3 rules: 1. kept if it was selected in at least seven datasets, 2. kept if selected in at least 8 datasets and 3. kept if selected in at least 9 datasets.

4. Finally, MissForest-Lasso was also run with 10 MissForest imputations instead of one imputation and coefficients estimates were averaged across imputed datasets for the final model (as for MICE-Lasso Musoro et al. 2014) in scenario S2. This was done in order to compare variable selection performance of the 10-imputation MissForest-Lasso with MICE-Lasso.

### 2.2.2 100-Covariate Dataset

The 100 covariate dataset was meant to reflect in a simplified way the real data at hand in terms of high dimensionality, different degrees of correlation between variables and number of levels of categorical variables.

The 100 variables were first drawn from a standard multivariate distribution with a fixed sample size of 500 and two different covariance (correlation) compound symmetry matrices: all variables weakly correlated with $\rho = 0.2$ or all variables highly correlated with $\rho = 0.8$. I did these choices of correlation because I wanted to study the case in which there was low correlation between variables and the case of multicollinearity, in contrast to the sparse mixed high and low correlations matrix used in the 20-covariates scenarios suggested by Musoro et al. (2014). The first 25 variables were subsequently dichotomised as follows:

- $\boldsymbol{X}_1$ to $\boldsymbol{X}_{10}$ at their 50th percentile
- $\boldsymbol{X}_{11}$ to $\boldsymbol{X}_{20}$ at their 30th percentile
- $\boldsymbol{X}_{21}$ to $\boldsymbol{X}_{25}$ at their 20th percentile

Finally, $X_{26}$ to $X_{50}$ were changed into categorical variables with 3 levels in the following way:

- $\boldsymbol{X}_{26}$ to $\boldsymbol{X}_{35}$ were split at their 20th and 80th percentiles
- $\boldsymbol{X}_{36}$ to $\boldsymbol{X}_{45}$ were split at their 30th and 60th percentiles
- $\boldsymbol{X}_{46}$ to $\boldsymbol{X}_{50}$ were split at their 40th and 70th percentiles

In the 100 covariate case, only one outcome generation was simulated, compared to the 20-covariate case, as dictated by the project real data moderation analysis needs: the true predictors of the outcome were 8 predictors (3 binary variables, 2 categorical with 3 levels and 3 continuous predictors) and 5 interaction terms between a binary predictor and 4 moderators (1 binary predictor, 1 polychoric predictor with 3 categories - i.e. 2 dummy variables - and 2 continuous predictors). The generative model was again a linear regression as in the 20 covariate scenario:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.21}$$

where $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{24}, \boldsymbol{X}_{26.1}, \boldsymbol{X}_{26.2}, \boldsymbol{X}_{27.1}, \boldsymbol{X}_{27.2}, \ldots, \boldsymbol{X}_{49.1}, \boldsymbol{X}_{49.2}, \boldsymbol{X}_{50}, \boldsymbol{X}_{51}, \ldots, \boldsymbol{X}_{100},$
$\boldsymbol{X}_1 \cdot \boldsymbol{X}_2, \boldsymbol{X}_1 \cdot \boldsymbol{X}_3, \ldots, \boldsymbol{X}_1 \cdot \boldsymbol{X}_{26.1}, \boldsymbol{X}_1 \cdot \boldsymbol{X}_{26.2}, \ldots, \boldsymbol{X}_1 \cdot \boldsymbol{X}_{100})$ was the design matrix of explanatory variables which columns were 249 covariates (included all the dummies, e.g. for the 3-levels categorical variable $\boldsymbol{X}_{26}$ the two dummies were $\boldsymbol{X}_{26.1}, \boldsymbol{X}_{26.2}$; and included all the interaction terms, no intercept), $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \boldsymbol{I})$ and $\boldsymbol{\beta}$ was the sparse 249-entry-vector of the coefficients, of which the non zero entries with respective covariates and dummy predictors were the following for a total of 15 true predictor terms:

- $\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3, \boldsymbol{X}_{26}, \boldsymbol{X}_{27}, \boldsymbol{X}_{51}, \boldsymbol{X}_{52}, \boldsymbol{X}_{53}$ with coefficients $\beta_1 = -0.7, \beta_2 = -0.8, \beta_3 = -0.9, \beta_{26.1} = -1.1, \beta_{26.2} = -1.3, \beta_{27.1} = 1.5, \beta_{27.2} = 1.2, \beta_{51} = 1,, \beta_{52} = 0.8, \beta_{53} = 0.7,$

- $\boldsymbol{X}_1 \cdot \boldsymbol{X}_2$, with $\beta_{1,2} = 0.7$,
- $\boldsymbol{X}_1 \cdot \boldsymbol{X}_{26.1}$, with $\beta_{1,26.1} = -1.1$,
- $\boldsymbol{X}_1 \cdot \boldsymbol{X}_{26.2}$, with $\beta_{1,26.2} = 1.3$,
- $\boldsymbol{X}_1 \cdot \boldsymbol{X}_{51}$, with $\beta_{1,51} = -0.9$,
- $\boldsymbol{X}_1 \cdot \boldsymbol{X}_{52}$, with $\beta_{1,52} = 0.8$.

In order to reflect the real data, all variables (included the outcome) but 6 predictors (3 TP and 3 FP) were set to have missing values MCAR and MAR in different scenarios, with high mean percentage of missing values (approximately 40%). In the case of MCAR data, observations to be missing were drawn from a binomial distribution with probabilities 0.2 for 27 variables ($\boldsymbol{X}_3$ to $\boldsymbol{X}_{10}, \boldsymbol{X}_{27}$ to $\boldsymbol{X}_{34}$ and from $\boldsymbol{X}_{53}$ to $\boldsymbol{X}_{63}$), 0.5 for other 67 variables ($\boldsymbol{X}_1, \boldsymbol{X}_{11}$ to $\boldsymbol{X}_{25}, \boldsymbol{X}_{35}$ to $\boldsymbol{X}_{50}$, $\boldsymbol{X}_{52}$, $\boldsymbol{X}_{64}$ to $\boldsymbol{X}_{97}$) and 0.2 for the outcome variable $\boldsymbol{Y}$. For MAR data, the probability of missingness depended on the 3 non-missing TP according to a logistic model per average percentage of missingness 20% and 50% for the same variables as for the MCAR case:

- covariates with 20% missingness had the missing probability given by $1+\exp(-3.1+X_2+0.9X_{26}-X_{51})$,
- covariates with 50% missingness had the missing probability given by $1+\exp(-0.5-X_2+X_{26}+0.1X_{51})$,
- the outcome had the 20% missing probability given by $1+\exp(-1-2.5X_2+2X_{51})$.

For the 100 covariate dataset scenario, Random Forests (RF) was run with conditional inference trees to take into account the correlation between variables and the different number of levels of the categorical variables (Strobl, Boulesteix, et al. 2008). The tuning parameter given by the number of variables chosen randomly at each split to build the trees was chosen as the one giving the lowest 'out-of-bag' (OOB) error.

The only scenarios considered for the 100 covariate design were the ones reflecting the project real data situation: S3 and S5 (see Table 2.2) with two degrees of correlation each (low correlation $\rho = 0.2$ and high correlation $\rho = 0.8$).

### 2.2.3 R packages, parallel computing and random number generators used

All the analyses were run in the statistical software R (Team 2016). This is a list of the R packages used in the simulation study:

- `doParallel` (Calaway et al. 2017): to parallelize computation
- `caret` (Kuhn 2016): implementing
  - Lasso and Elasticnet from `glmnet` (Friedman, Hastie, and Tibshirani 2010)
  - Random Forests from `randomForest` (Liaw and Wiener 2002a)
  - Conditional inference Random Forests from `party` (Hothorn, B uhlmann, et al. 2006)
- `mice` (Van Buuren and Groothuis-Oudshoorn 2011): to run multivariate imputation using chained equations
- `missForest` (Stekhoven 2013): to run the algorithm of MissForest (iterative non-parametric missing value imputation through Random Forests)

Parallel computing was performed on an Intel Haswell E5-2660 v3 machine with 2.60GHz of frequency and 10 cores (8GB ram per core).

I used the R random number generator by specifying the seed at each step between and within simulations which required it, and making sure that the sequence of random numbers was long before repetition in the case of the 20 covariate dataset simulations (see Subsection 2.2.1). In the 100 covariate study (subsection 2.2.2), I only fixed the starting seeds. In the case of parallel computing, I used the command `clusterSetRNGStream` in the package `doParallel` to fix the seeds (see R code in the Appendix section B.2).

### 2.2.4 Encountered problems

There were many problems I experienced in order to run this simulation study. Firstly, as the considered methods algorithms required a large amount of computational time to run, I learnt parallel computing to speed up the jobs and I familiarised with the operating-system Linux in order to use Google Cloud, a Goldsmith server and the King's College Rosalind cluster, that were the only facilities allowing me to run these computationally expensive simulation jobs. Without a strong background in computer science, I learnt how to write more efficient R code to further diminish the simulation running time. This learning process took me months of trial and error. I started by running R in Google Cloud which had more memory and more parallel cores options, but it was too expensive. Thus, I was given an account in a dssc server by Dr. Daniel Stamate (Department of Computing at Goldsmith University). Even though I had enough memory, I could only run one job on multiple cores at a time. Therefore, I used the Rosalind cluster to run multiple batch jobs on multiple cores each with the required memory

to increase efficiency of simulations. I learnt some Linux language and how to set a batch job and different ways to save the jobs outputs. I ended up to improve the timing of each set of simulations to 20 days, instead of 45 days, being able to run at least 10 sets of simulations at the same time depending on the cluster availability. The MissForest-Random Forests and MICE-Lasso/Elasticnet simulations were the most computationally expensive in time: MICE because of the 10 imputations and Random Forests because of its bootstrap tuning procedure. However, apart from the queues to wait before a job started to run, the cluster has had plenty of failures both in the parallel environment and in the cluster storage by causing my simulation study progress to slow down significantly.

Secondly, after having started using the Rosalind cluster and having got the first results, I realized that there was an error in the R code used by Musoro et al. 2014 to run their simulations, which I replicated (see the error in the Appendix section B.1). Thus, I had to edit the error and rerun the jobs in order to get the correct results, which therefore differed in parts from Musoro et al. (2014). I wrote to the authors about their incorrect code without receiving any answer.

## 2.3 Simulation results

### 2.3.1 Results from 20-covariate datasets, 10 true predictors

#### MICE-Lasso and MICE-Elasticnet: 20-covariate data results

**Lasso S1: No missing data, No assumption of moderation** In absence of missing observations and interaction terms in the true linear predictor, when the sample size was 250, the Lasso best apparent model (with bootstrap tuning) on average overfitted the data, by selecting 52.9% (SD 18.7) of the false predictors (FP) altogether with 99.8% (SD 1.4) of the true predictors (TP) for an acceptable PPV of 66.4% (SD 8.7) (see Table 2.3). As a consequence, only 0.7% of the times the Lasso selected the true model covariates and 1.7% of the times it selected the true model but only one TP. Moreover, only 58.7% of the times the 10 TPs were the top 10 ranked selected variables, while 9 of them were in the top 10 variables for 95.0% of the times. Overfitting was reflected in the estimated average apparent MSE (2.840 with 2.5th and 97.5th percentiles: 2.315 and 3.406) which was lower than theoretical MSE (3.028) and subsequently much lower than the bootstrap-corrected MSE (3.262, 2.5th and 97.5th percentiles: 2.663 and 3.946, see Table 2.4), giving an average pseudo-$R^2$ of 0.669 (2.5th and 97.5th percentiles being 0.606 and 0.732, the outcome mean variance of the simulated datasets was

10.539 (SD 0.954)). The estimate of internal optimism for the MSE was slightly higher in absolute value than the estimate of external optimism showing that the resampling procedure gave nearly unbiased estimates of optimism. However, the discrepancy between internal and external estimates of optimism got bigger with the increasing tolerance percentage of the tuning parameter. In concordance with Musoro et al. 2014 results, there was suboptimal calibration due to overshrinkage, showed by the mean calibration slope $\beta_{LP}$ which was higher than 1 (see Table 2.4).

The Lasso 3% tolerance models performed on average better in terms of variable selection, at the cost of less but still acceptable accuracy in prediction ($MSE_{corrected}$=3.474, 2.5th and 97.5th percentiles 2.834 and 4.215, see Table 2.4). Instead, the 1 SE tolerance models were similar to the best models and the 15% tolerance models were too much parsimonious with poor prediction accuracy (see Table 2.4). The 3% tolerance models on average selected 98.0% (SD 4.0) of the TPs together with only 15.6% (SD 12.5) of the FPs for a good PPV of 87.3% (SD 9.0). Also, the number of times in which the 10 TPs were the top ten selected variables in terms of absolute value and the number of times the true model was selected were larger than those for the best model (see Table 2.3).

The 1000 observation datasets models gave improved and more precise accuracy estimates (the best model $MSE_{corrected}$ was 3.092 (2.819 to 3.397) and relative pseudo-$R^2$ was 0.704 (0.645 to 0.758), the mean outcome variance of simulated datasets being 10.552 (SD 0.461), see Table A.3 in the Appendix for the other measures) and performed slightly better variable selection (see Table 2.3). Both internal and external MSE optimism estimates were greatly reduced in absolute value and resulted much closer to each other, so that the bias due to resampling was now lower.

**Minor analyses 1-2 results** The simulations for the Lasso model, run with 20 times repeated 10-cross-validation tuning instead of bootstrap did not show any improved performance in terms of overfitting and the computational time required was larger: it took approximately 3 days longer then the analysis with bootstrap tuning.

The simulations of Lasso in which the bootstrap validation was done by testing the bootstrap models on the OOB data obtained average test MSEs (see tables A.1 and A.2 in the Appendix) similar to the average MSE bootstrap-corrected estimate from Harrell's bootstrap procedure (see tables 2.4 and A.3), but the increased computational time made us entirely opt for Harrell's bootstrap validation methods (Harrell, Lee, and Mark 1996).

Table 2.3: **Variable selection** simulation study results for scenarios **S1** (without missing data, no assumption of moderation) and **S2** (with missing data, complete outcome, no assumption of moderation) for **Lasso** and **MICE-Lasso** best and tolerance models in the case of **20 covariates** and 300 samples of 250 and 1000 observations. The following results are shown: the estimated percentages of the times all the 10 true predictors (TP) are selected at the same time, the estimated percentages of the times the true model is selected apart from one variable, the percentages of the times the TP are the top ranked variables among the selected, the percentage of times the TP are the top ranked variables apart from one TP, the average percentages of selected TP among the TP (sensitivity, SEN), the average percentages of selected false positive predictors (FP) among FP (false positive rate, FPR), and the average percentages of selected TP among the selected variables (positive predictive value, PPV). The mean percentages are shown along with their standard deviation (SD).

| | **LASSO** | | | | **MICE-LASSO** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Variable selection | complete | | | | MCAR | | | | MAR | | | |
| | Best | 1 SE tolerance | 3% tolerance | 15% tolerance | Best | 1 SE tolerance | 3% tolerance | 15% tolerance | Best | 1 SE tolerance | 3% tolerance | 15% tolerance |
| | | | | | | **250 observations** | | | | | | |
| **% true models** | 0.7 | 3.7 | 16.0 | 13.0 | 0 | 0 | 0 | 1.7 | 0 | 0 | 0 | 2.7 |
| **% true models but 1 P** | 1.7 | 11.0 | 32.0 | 47.7 | 0 | 0 | 0 | 13.0 | 0 | 0 | 0.7 | 19.3 |
| **% TP in top 10** | 58.7 | 62.3 | 60.0 | 14.7 | 21.3 | 30.7 | 38.0 | 19.0 | 30.0 | 38.0 | 42.3 | 21.0 |
| **% TP in top 10 but 1** | 36.3 | 33.0 | 33.7 | 51.7 | 44.7 | 44.3 | 41.0 | 44.3 | 39.7 | 40.3 | 39.0 | 41.3 |
| **SEN (SD)** | 99.8 (1.4) | 99.5 (2.25) | 98.0 (4.0) | 87.6 (8.9) | 100 (0) | 99.9 (0.8) | 99.5 (2.3) | 93.3 (7.8) | 100 (0) | 99.8 (1.3) | 99.3 (2.7) | 92.9 (7.4) |
| **FPR (SD)** | 52.9 (18.7) | 33.4 (17.6) | 15.6 (12.5) | 1.8 (4.6) | 99.4 (3.0) | 94.6 (9.2) | 74.1 (17.7) | 22.7 (15.2) | 99.4 (2.8) | 91.6 (11.08) | 66.4 (18.7) | 18.6(13.8) |
| **PPV (SD)** | 66.4 (8.7) | 76.2 (10.1) | 87.3 (9.0) | 98.3 (4.4) | 50.2 (0.8) | 51.5 (2.7) | 58.0 (6.3) | 81.8 (9.9) | 50.2 (0.7) | 52.3 (3.4) | 60.7 (7.3) | 84.6 (9.8) |
| | | | | | | **1000 observations** | | | | | | |
| **% true models** | 0.3 | 7.3 | 80.3 | 18.3 | 0 | 1.7 | 25.3 | 4.0 | 0 | 1.7 | 25.3 | 33.7 |
| **% true models but 1 P** | 1.0 | 19.3 | 17.7 | 81.7 | 0 | 9.0 | 36.0 | 18 | 0 | 0 | 15.7 | 57.0 |
| **% TP in top 10** | 99.3 | 99.7 | 99.0 | 18.3 | 92.3 | 97.3 | 65.0 | 22.3 | 95.0 | 98.0 | 97.0 | 36.7 |
| **% TP in top 10 but 1** | 0.7 | 0.3 | 1.0 | 81.7 | 7.7 | 2.7 | 31.7 | 16.7 | 5.0 | 2.0 | 3.0 | 63.3 |
| **SEN (SD)** | 100 (0) | 100 (0) | 99.9 (1.0) | 91.9 (3.9) | 100 (0) | 100 (0) | 100.0 (0.6) | 94.1 (4.9) | 100 (0) | 100 (0) | 100 (0) | 93.9 (4.9) |
| **FPR (SD)** | 51.3 (19.1) | 25.2 (15.6) | 2.1 (4.5) | 0.1 (0.6) | 99.7 (2.4) | 88.2 (11.6) | 32.8 (15.7) | 2.4 (4.7) | 99.5 (2.4) | 86.4 (12.0) | 27.8 (15.1) | 1.5 (3.9) |
| **PPV (SD)** | 67.2 (8.6) | 81.0 (9.7) | 98.2 (4.0) | 99.9 (0.5) | 50.1 (0.7) | 53.3 (3.5) | 76.3 (9.0) | 97.8 (4.4) | 50.1 (0.6) | 53.9 (3.7) | 79.3 (9.3) | 98.5 (3.7) |

Table 2.4: **Accuracy** simulation study results for **MICE-Lasso** analysis with Harrell (1996) bootstrap validation: scenarios **S1** (without missing data, no assumption of moderation) and **S2** (with missing data, complete outcome, no assumption of moderation) based on 300 data sets of **20 variables** each (**n=250**). Means of all estimates along with their corresponding 2.5th and 97.5th percentile values within parenthesis. The theoretical MSE is $1.74^2 = 3.028$

| **Estimates** | **Complete cases** | | | |
|---|---|---|---|---|
| | Best model | 1 SE tolerance | 3% tolerance | 15% tolerance |
| $MSE_{apparent}$ | 2.840 (2.315,3.406) | 2.937 (2.398,3.544) | 3.127 (2.553,3.772) | 4.002 (3.271,4.905) |
| $\beta_{LP}$ | 1.062 (1.041,1.084) | 1.102 (1.075,1.131) | 1.161 (1.124,1.200) | 1.334 (1.261,1.429) |
| Tuning $\lambda$ | 0.065 (0.039,0.085) | 0.106 (0.076,0.134) | 0.168 (0.134,0.209) | 0.360 (0.293,0.410) |
| $MSE_{ext}$ | 3.509 (3.289,3.791) | 3.590 (3.347,3.890) | 3.781 (3.486, 3.172) | 4.718 (4.113,5.420) |
| $Optimism_{ext}$ | -0.669 (-1.212,-0.010) | -0.653 (-1.198,-0.014) | -0.654 (-1.216,0.055) | -0.716 (-1.460,0.176) |
| $Optimism_{int}$ | -0.422 (-0.534,-0.332) | -0.383 (-0.482,-0.295) | -0.348 (-0.442,-0.267) | -0.294 (-0.404,-0.216) |
| $MSE_{corrected}$ | 3.262 (2.663,3.946) | 3.320 (2.716,4.034) | 3.474 (2.834,4.215) | 4.311 (3.534,5.291) |
| $\beta_{LP^*}$ | 1.028 (1.015,1.042) | 1.069 (1.051,1.088) | 1.122 (1.100,1.157) | 1.309 (1.235,1.379) |
| | | **MCAR** | | |
| $MSE_{apparent}$ | 2.842 (2.204,3.622) | 2.916 (2.250,3.726) | 3.094 (2.379,3.952) | 3.955 (2.999,5.156) |
| $\beta_{LP}$ | 1.038 (1.014,1.063) | 1.078 (1.048,1.108) | 1.128 (1.095,1.179) | 1.325 (1.233,1.418) |
| Tuning $\lambda$ | 0.046 (0.024,0.074) | 0.085 (0.053,0.122) | 0.145 (0.097,0.201) | 0.354 (0.244,0.466) |
| $MSE_{ext}$ | 3.683 (3.343,4.209) | 3.700 (3.344,4.211) | 3.829 (3.446,4.439) | 4.676 (3.919,5.642) |
| $Optimism_{ext}$ | -0.841 (-1.736,0.001) | -0.784 (-1.632,0.054) | -0.735 (-1.528,0.147) | -0.721 (-1.493,0.253) |
| $Optimism_{int}$ | -0.848 (-1.157,-0.590) | -0.743 (-1.015,-0.508) | -0.647 (-0.860,-0.415) | -0.428 (-0.613,-0.294) |
| $MSE_{corrected}$ | 3.690 (2.880,4.650) | 3.659 (2.840,4.643) | 3.709 (2.865,4.698) | 4.478 (3.347,5.589) |
| $\beta_{LP^*}$ | 0.959 (0.926,0.988) | 1.001 (0.965,1.033) | 1.044 (1.017,1.097) | 1.229 (1.161,1.312) |
| | | **MAR** | | |
| $MSE_{apparent}$ | 2.868 (2.280,3.601) | 2.951 (2.335,3.712) | 3.138 (2.470,3.939) | 4.026 (3.106,5.097) |
| $\beta_{LP}$ | 1.045 (1.022,1.070) | 1.085 (1.058,1.119) | 1.145 (1.105,1.197) | 1.323 (1.243,1.437) |
| Tuning $\lambda$ | 0.049 (0.026,0.072) | 0.089 (0.058,0.120) | 0.150 (0.106,0.193) | 0.347 (0.261,0.449) |
| $MSE_{ext}$ | 3.644 (3.296,4.112) | 3.683 (3.338,4.119) | 3.836 (3.428,4.312) | 4.740 (4.041,5.590) |
| $Optimism_{ext}$ | -0.776 (-1.618,0.039) | -0.732 (-1.501,0.053) | -0.699 (-1.417,0.080) | -0.711 (-1.556,0.218) |
| $Optimism_{int}$ | -0.714 (-0.942,-0.523) | -0.631 (-0.835,-0.466) | -0.563 (-0.721,-0.399) | -0.394 (-0.551,-0.275) |
| $MSE_{corrected}$ | 3.582 (2.874,4.471) | 3.582 (2.859,4.479) | 3.681 (3.474,5.621) | 4.434 (3.474,5.621) |
| $\beta_{LP^*}$ | 0.982 (0.956,1.005) | 1.025 (0.994,1.051) | 1.068 (1.045,1.122) | 1.257 (1.192,1.344) |

Table 2.5: **Variable selection** simulation study results for scenarios **S1** (without missing data, no assumption of moderation) and **S2** (with missing data, complete outcome, no assumption of moderation) for **MICE-Elasticnet** best and tolerance models in the case of **20 covariates** and 300 samples of 250 and 1000 observations. The following results are shown: the estimated percentages of the times all the 10 true predictors (TP) are selected at the same time, the estimated percentages of the times the true model is selected apart from one variable, the percentages of the times the TP are the top ranked variables among the selected, the percentage of times the TP are the top ranked variables apart from one TP, the average percentages of selected TP among the TP (sensitivity, SEN), the average percentages of selected false positive predictors (FP) among FP (false positive rate, FPR), and the average percentages of selected TP among the selected variables (positive predictive value, PPV). The mean percentages are shown along with their standard deviation (SD)

| | ELASTICNET | | | | MICE-ELASTICNET | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Variable selection** | **Complete data** | | | | **MCAR** | | | | **MAR** | | | |
| | Best | 1 SE tolerance | 3% tolerance | 15% tolerance | Best | 1 SE tolerance | 3% tolerance | 15 %tolerance | Best | 1 SE tolerance | 3% tolerance | 15% tolerance |
| | | | | | | **250 observations** | | | | | | |
| **% true models** | 0 | 19.0 | 5.3 | 0.7 | 0 | 0 | 0.3 | 0 | 0 | 0.3 | 0.3 | 1.0 |
| **% true models but 1 P** | 2.0 | 33.0 | 18.7 | 4.0 | 0 | 0 | 0.3 | 3 | 0 | 1.0 | 2.0 | 5.7 |
| **% TP in top 10** | 56.7 | 44.7 | 15.3 | 2.7 | 20.7 | 31.3 | 34.0 | 27.7 | 28.3 | 35.3 | 31.0 | 20.0 |
| **% TP in top 10 but 1** | 36.3 | 44.3 | 28.0 | 7.7 | 43.3 | 42.7 | 40.0 | 36.3 | 37.7 | 41.7 | 47.7 | 38.3 |
| **SEN (SD)** | 99.8 (1.4) | 96.0 (6.0) | 80.3 (18.2) | 36.0 (30.1) | 100 (0) | 99.7 (1.6) | 99.3 (3.0) | 94.9 (13.5) | 100 (0) | 99.6 (2.1) | 98.8 (3.9) | 92.1 (18.2) |
| **FPR (SD)** | 57.0 (20.6) | 18.6 (25.2) | 10.2 (20.8) | 3.6 (10.9) | 99.8 (1.3) | 94.1 (14.3) | 84.1 (23.5) | 52.2 (26.5) | 99.7 (1.8) | 90.0 (19.0) | 77.4 (28.8) | 45.5 (28.3) |
| **PPV (SD)** | 64.8 (9.0) | 87.2 (14.7) | 93.0 (12.7) | 97.2 (7.9) | 50.0 (0.3) | 51.8 (4.9) | 55.4 (9.4) | 67.6 (13.3) | 50.1 (0.5) | 53.3 (7.2) | 58.2 (12.3) | 71.4 (15.3) |
| | | | | | | **1000 observations** | | | | | | |
| **% true models** | 0.3 | 83.3 | 0.7 | 0.3 | 0 | 1.7 | 25.3 | 4.0 | 0 | 3.0 | 26.0 | 3.0 |
| **% true models but 1 P** | 0.7 | 14.3 | 33.3 | 0.3 | 0 | 9.0 | 36.0 | 18.0 | 0 | 13.7 | 46.7 | 13.3 |
| **% TP in top 10** | 99.3 | 98.7 | 1.3 | 0.3 | 92.3 | 97.3 | 65.0 | 22.3 | 95.0 | 97.0 | 56.3 | 14.7 |
| **% TP in top 10 but 1** | 0.7 | 1.3 | 33.0 | 0.3 | 7.7 | 2.7 | 31.7 | 16.7 | 5.0 | 3.0 | 40.3 | 12.0 |
| **SEN (SD)** | 100 (0) | 99.9 (1.0) | 79.9 (9.5) | 14.5 (8.7) | 100 (0) | 100.0 (0.6) | 96.4 (5.5) | 60.6 (33.4) | 100 (0) | 100 (0) | 95.6 (5.8) | 52.4 (32.5) |
| **FPR (SD)** | 52.0 (19.0) | 2.2 (7.4) | 0.1 (1.8) | 0 (0) | 99.6 (2.5) | 57.5 (31.3) | 20.5 (30.9) | 6.4 (13.8) | 99.4 (2.5) | 46.3 (30.1) | 13.2 (24.8) | 3.6 (10.9) |
| **SEN (SD)** | 66.8 (8.5) | 98.18 (5.1) | 99.9 (1.4) | 100 (0) | 50.1 (0.7) | 66.2 (14.0) | 87.3 (17.3) | 95.2 (9.7) | 50.2 (0.7) | 71.2 (14.2) | 91.4 (14.4) | 97.3 (7.7) |

Table 2.6: **Accuracy** simulation study results for **MICE-Elasticnet** analysis with Harrell bootstrap validation: scenarios **S1** (without missing data, no assumption of moderation) and **S2** (with missing data, complete outcome, no assumption of moderation) based on 300 data sets of **20 variables** each (**n=250**). Means of all estimates along with their corresponding 2.5th and 97.5th percentile values within parenthesis. The theoretical MSE is $1.74^2 = 3.028$

| **Estimates** | **Complete cases** | | | |
|---|---|---|---|---|
| | Best model | 1 SE tolerance | 3% tolerance | 15% tolerance |
| $MSE_{apparent}$ | 2.839 (2.318,3.408) | 3.360 (2.568,4.215) | 4.451 (2.836,6.060) | 6.858 (3.683,9.430) |
| $\beta_{LP}$ | 1.069 (1.050,1.090) | 1.232 (1.118,1.342) | 1.433 (1.197,1.680) | 2.670 (1.464,6.064) |
| Tuning $\alpha$ | 0.605 (0.100,0.900) | 0.900 (0.900,0.900) | 0.900 (0.900,0.900) | 0.900 (0.900,0.900) |
| Tuning $\lambda$ | 0.114 (0.065,0.251) | 0.352 (0.220,0.455) | 0.667 (0.455,0.739) | 1.641 (1.199,1.947) |
| $MSE_{ext}$ | 3.519 (3.282,3.818) | 4.056 (3.508,4.756) | 5.154 (3.681,6.548) | 7.472 (4.496,9.775) |
| $Optimism_{ext}$ | -0.678 (-1.238,-0.020) | -0.696 (-1.316,0.050) | -0.702 (-1.480,0.207) | -0.614 (-1.783,0.779) |
| $Optimism_{int}$ | -0.848 (-1.157,-0.590) | -0.743 (-1.015,-0.508) | -0.626 (-0.860,-0.415) | -0.445 (-0.613,-0.294) |
| $MSE_{corrected}$ | 3.687 (3.005,4.489) | 4.103 (3.215,5.092) | 5.077 (3.429,6.759) | 7.302 (4.169,9.969) |
| $\beta_{LP^*}$ | 0.959 (0.926,0.988) | 1.001 (0.965,1.033) | 1.060 (1.017,1.097) | 1.233 (1.161,1.312) |
| | **MCAR** | | | |
| $MSE_{apparent}$ | 2.843 (2.203,3.622) | 3.067 (2.294,3.993) | 3.556 (2.515,4.952) | 5.222 (3.370,7.734) |
| $\beta_{LP}$ | 1.049 (1.024,1.075) | 1.157 (1.075,1.263) | 1.303 (1.153,1.511) | 1.893 (1.407,3.091) |
| Tuning $\alpha$ | 0.370 (0.110,0.785) | 0.899 (0.880,0.900) | 0.900 (0.900,0.900) | 0.900 (0.900,0.900) |
| Tuning $\lambda$ | 0.125 (0.061,0.209) | 0.316 (0.194,0.450) | 0.585 (0.385,0.820) | 1.517 (1.066,2.060) |
| $MSE_{ext}$ | 3.691 (3.345,4.209) | 3.856 (3.441,4.487) | 4.319 (3.642,5.686) | 5.919 (4.377,8.230) |
| $Optimism_{ext}$ | -0.848 (-1.731,-0.011) | -0.789 (-1.563,0.088) | -0.763 (-1.497,0.201) | -0.697 (-1.697,0.477) |
| $Optimism_{int}$ | -0.848 (-0.934,-0.752) | -0.743 (-0.825,-0.654) | -0.647 (-0.715,-0.566) | -0.428 (-0.480,-0.374) |
| $MSE_{corrected}$ | 3.690 (3.343,4.010) | 3.659 (3.324,3.968) | 3.709 (3.362,4.021) | 4.478 (4.040,4.859) |
| $\beta_{LP^*}$ | 0.959 (0.949,0.971) | 1.001 (0.990,1.012) | 1.044 (1.031,1.057) | 1.229 (1.208,1.250) |
| | **MAR** | | | |
| $MSE_{apparent}$ | 2.869 (2.280,3.602) | 3.142 (2.394,4.110) | 3.709 (2.609,5.227) | 5.485 (3.493,8.123) |
| $\beta_{LP}$ | 1.056 (1.032,1.082) | 1.175 (1.092,1.278) | 1.332 (1.169,1.553) | 1.996 (1.444,3.307) |
| Tuning $\alpha$ | 0.403 (0.100,0.810) | 0.899 (0.899,0.900) | 0.900 (0.900,0.900) | 0.900 (0.900,0.900) |
| Tuning $\lambda$ | 0.128 (0.067,0.224) | 0.330 (0.210,0.451) | 0.608 (0.417,0.820) | 1.558 (1.108,2.066) |
| $MSE_{ext}$ | 3.655 (3.317,4.122) | 3.893 (3.473,4.512) | 4.454 (3.615,5.749) | 6.187 (4.404,8.540) |
| $Optimism_{ext}$ | -0.786 (-1.613,0.026) | -0.751 (-1.491,0.031) | -0.746 (-1.477,0.092) | -0.702 (-1.715,0.386) |
| $Optimism_{int}$ | -0.706 (-0.931,-0.523) | -0.589 (-0.778,-0.433) | -0.496 (-0.659,-0.349) | -0.351 (-0.489,-0.216) |
| $MSE_{corrected}$ | 3.576 (2.865,4.461) | 3.731 (2.888,4.723) | 4.205 (3.050,5.744) | 5.837 (3.806,8.477) |
| $\beta_{LP^*}$ | 0.995 (0.965,1.017) | 1.084 (1.034,1.137) | 1.204 (1.124,1.306) | 1.693 (1.415,2.152) |

**Elasticnet S1: No missing data, No assumption of moderation** The Elasticnet model underperformed the Lasso by overfitting more the data in the best model case (see Table 2.6, Lu and Petkova 2014). However, all the results were similar to the Lasso results (the Elasticnet tuning parameter $\alpha$ for the best model was always close to 1 for the best model) apart from slightly higher estimates of the MSE optimism in absolute value and higher estimates of corrected MSE for the best models. The shrinkage got stronger with increasing model tolerance compared to the Lasso model. Only the 1 SE tolerance model could give acceptable variable selection as the other stronger penalties led to drastic shrinkage of the coefficients and very poor accuracy. The internal and external MSE optimism estimates were closer to each other compared to the Lasso estimates in the case of 250 observations, revealing that Elasticnet internally validated performance was likely to be more similar to the performance on new data than the Lasso.

In the 1000 observation dataset models the external optimism estimates were very similar to the internal optimism estimates apart from the 3% and 15% tolerance model corresponding estimates which diverted from each other (see Table A.4 in the Appendix). This indicates that the estimates of internal optimism were biased for stronger penalties. While Elasticnet variable selection improved in the best models and 1SE models when the sample size was 1000, it worsened for higher model tolerance levels as in these cases Elasticnet over-shrank the coefficients and also excluded the TPs (see Figure 2.27).

**MICE-Lasso S2: Missing data, No assumption of moderation, complete outcome** When missing data were present, the MICE-Lasso best model (Musoro et al. 2014) selected all the variables, both false and true predictors, 98.3% of the times (see Figures 2.29 and 2.28), giving a poor variable selection performance (see Table 2.3). Only the models with MSE within 3% or 15% of the minimum had acceptable variable selection. The noise variables with more missing data ($\boldsymbol{X}_3$ and $\boldsymbol{X}_{13}$, 50% missingness, see Subsection 2.2.1), not correlated with any other variable, were selected almost 100% of the times, also in the 3% tolerance model. Moreover, the best and 1SE tolerance models never selected the true model nor the true model but one TP and the larger tolerance models only chose all the TPs but one TP up to 19.3% of the times in the case of MAR data and up to 13% of the times for the MCAR data (see Table 2.3).

The best model variable selection result was unexpected, because Musoro et al (2014) found a much better result: the best model still retained a large number of irrelevant covariates but with a much lower selection frequency ranging from about 66% to 75%. The authors explained the large quantity of chosen FPs, as one may expect, with the fact that covariates were

counted every time they were selected by Lasso in at least one of the imputed datasets. The big difference between my result and Musoro's result is due to an error in Musoro's function to compute the best and tolerance models: their best $\lambda$ was in fact a tolerance $\lambda$ and their given tolerance $\lambda$s were actually even stronger penalties (see Appendix section B.1).

Even though the method performed poorly in variable selection, the prediction accuracy result for the 250 observations scenario was similar compared to the complete data case and within 22% of the theoretical MSE. There was a decrease of only approximately 6% in the corrected-pseudo-$R^2$ for the best model (optimism-corrected MSE being 3.690, 2.5th and 97.5th percentiles: 2.880 and 4.650, for MCAR data and 3.582, 2.5th and 97.5th percentiles: 2.874 and 4.471, for MAR data, see Table 2.4) compared to the complete data scenario. The increased penalty MSEs were still acceptable apart from the 15% tolerance MSE which was very poor. Calibration performance slightly improved compared to scenario S1 without missing data. The combined model dealt with MCAR data slightly worse than MAR data.

When the sample size was larger (1000 observations), the increase in the corrected-pseudo-$R^2$ was only about 1% compared to the smaller sample size and all estimates were more precise.

**Minor analysis 3 results** The results from the 'majority method' selection (see the Methods Subsection 2.2.1) did not improve variable selection.

**MICE-Elasticnet S2: Missing data, No assumption of moderation, complete outcome** MICE-Elasticnet performance was similar to MICE-Lasso in both prediction accuracy and variable selection apart from the fact that given the same tolerance level, MICE-Elasticnet gave more parsimonious models in the case of 1000 observations (see tables 2.5, 2.6 and A.4, last one in the Appendix). MICE-Elasticnet performed better than Elasticnet (scenario S1 without missing data) in terms of prediction accuracy (see Figures 2.8, A.5, 2.11, A.8 ).

**Lasso S3: No missing data, Assumption of moderation** When interaction terms were present among the predictors and there were no missing data, the average optimism-corrected MSE for the best Lasso model was acceptable compared to the theoretical MSE: 3.410 (2.5th and 97.5th percentiles: 2.788 and 4.131, see Table 2.9) for the 250 observation datasets, giving a corrected-pseudo-$R^2$ of 0.774 (percentiles being 0.722 and 0.820, the mean variance of the outcome across the simulated datasets being 15.150 (SD 1.385)). Corrected MSE corresponding to stronger penalties were still within 20% of the theoretical MSE apart from the 15%

tolerance model MSE which was very poor. Overall, the average estimated internal optimism for the MSE was larger in absolute value than the optimism for the case without interactions showing that having double the number of covariates in the model (40) with the same number of people (250) increases the bias due to overfitting. Calibration slope estimates were just above 1, similar in value to the ones corresponding to scenario S1 in which interactions were not included.

Variable selection performance for the Lasso was slightly inferior to the performance in the case without interactions (see Table 2.7). On average, 92.3% (SD 3.3) TPs among the actual TPs were selected in the best model for an acceptable PPV of 62.6% (SD 7.9). The PPV increased to 90.4% (SD 4.4) with the 15% tolerance model, accompanied by a small decrease in the percentage of selected TPs among the actual TPs, for a very good variable selection. All the tolerance models performed better than the best model in terms of variable selection, but the 15% tolerance model had poor prediction accuracy (see Table 2.9). Another important fact is that both best and tolerance models hardly ever selected exactly the TPs in the model, they always included FPs (see Table 2.7). Instead, scenario S1 without interactions showed that at least the 3% and 15% tolerance model applied to 32% and 47.7% of the datasets respectively selected exactly the true model predictors but one TP (2.3). By looking at figure 2.30, I can see that in particular one true predictor was hardly ever selected (V1) as its true coefficient is the smallest and it strongly correlates with V5, a noise variable ($\rho_{1,5} = 0.72$, see Subsection 2.2.1).

In the case of 1000 observations, all results improved in both prediction accuracy and variable selection (see tables 2.7 and A.5 in the Appendix).

**Elasticnet S3: No missing data, Assumption of moderation** Elasticnet behaved like the Lasso in presence of interaction terms among the predictors (see tables 2.8, 2.10 and A.6, last table in the Appendix). Again, given the same tolerance level, Elasticnet tended to penalise more than Lasso by resulting in poorer prediction accuracy compared to Lasso tolerance models.

**MICE-Lasso S4: Missing data, Assumption of moderation, complete outcome** When interactions were in the linear predictor and there were missing data, the prediction accuracy of MICE-Lasso was inferior compared to the accuracy obtained when missing data were absent (scenario S3). The missing data imputation through MICE caused the model to have a higher optimism-corrected MSE for the best model, which was poor: 4.119 (2.5th and 97.5th percentiles: 3.168 and 5.157, see Table 2.9), still giving a good corrected pseudo-$R^2$ of 0.727

(2.5th and 97.5th percentiles: 0.656 and 0.791) for the 250 observations dataset with MCAR data. MAR data gave slightly better results. The bias due to optimism was greatly reduced when the sample size was 1000 (see Table A.5).

The variable selection performance was as bad as scenario S2: the best model selected all the variables almost 100% of the times, irrespective of them being noise variables or true predictors (see Table 2.7 and figures 2.31 and 2.33). Again the variables $\boldsymbol{X}_3$ and $\boldsymbol{X}_{13}$ (noise variables with a large percentage of missing data, see Subsection 2.2.1) were always selected because of the filled in values by MICE that generated correlation between the imputed variables and the outcome.

**MICE-Elasticnet S4: Missing data, Assumption of moderation, complete outcome** The performances for MICE-Elasticnet were again very similar to MICE-Lasso for this scenario (see tables 2.8, 2.10 and A.6, last table in the Appendix). Another time MICE-Elasticnet outperformed Elasticnet (scenario S3 without missing data) because of MICE imputation of missing data, but this time only in the accuracy performance of the tolerance models with tolerance levels 3% or higher.

**MICE-Lasso S5: Assumption of moderation, Missing data also in outcome (20% missingness MAR and MCAR)** Missingness in the outcome affected negatively the accuracy performance of MICE-Lasso, which was already poor, by decreasing the optimism-corrected pseudo-$R^2$ (calculated using the variance of the complete outcome) by approximately 4% in the case of MCAR data (optimism-corrected MSE: 4.576, 2.5th and 97.5th percentiles being 3.387 and 5.892) and by 2% in the case of MAR data (optimism-corrected MSE: 4.385, 2.5th and 97.5th percentiles being 3.320 and 5.644) compared to scenario S4 (without missing data in the outcome) for the 250 observation datasets (see Table 2.13). There was a consistent difference between MAR and MCAR results (see also scenario S2 for MICE-Lasso above) favouring MAR results. Still good proportions of variance were explained by the model applied to both missing data types.

As regards variable selection, MICE-Lasso tended to include all variables in the model also in this scenario. The variable selection performance was worse than in scenario S4 (see Table 2.11).

The larger sample size case did provide better results (see tables A.7 and 2.11, however this scenario performance was inferior to the other scenarios.

**MICE-Elasticnet S5: Assumption of moderation, Missing data also in outcome (20% missingness MAR and MCAR)** Performance for MICE-Elasticnet and MICE-Lasso were close apart from the fact that MICE-Elasticnet penalised more the models given the same tolerance level (see tables 2.12, 2.14 and A.8, last table in the Appendix).

**MICE-Lasso S6: Interactions in the imputation model, complete outcome** Adding interaction terms in the imputation model when the outcome was complete did not improve the prediction accuracy of MICE-Lasso. The average internal optimism estimates were the largest amongst the MICE-Lasso scenarios (best model: -1.451, 2.5th and 97.5th percentiles -1.965 and -1.047, see Table 2.16 and figure 2.4) for 250 observation datasets. Also the difference between internal and external MSE optimism was the worst, meaning that adding the 20 interactions terms in the imputation model caused more noise in the missing data filling, as only four interaction terms were TP (see Subsection 2.2.1).

The variable selection performance was also inferior to the other scenarios for MICE-Lasso for MCAR data, while it was almost equivalent to scenario S5 for MAR data (see Table 2.15 and figures 2.6, A.1 and 2.7, last one with MAR data in the Appendix).

The 1000 observation datasets results were improved with respect to the smaller sample size data (see tables 2.15 and A.9 and figures 2.2, 2.4, 2.6, A.3 and A.2, last two figures in the Appendix).

Table 2.7: **Variable selection** simulation study results for scenarios **S3** (assumption of moderation, without missing data) and **S4** (assumption of moderation, with missing data) for **MICE-Lasso** best and tolerance models in the case of **20 covariates** and 300 samples of 250 and 1000 observations. The following results are shown: the estimated percentages of the times all the 10 true predictors (TP) are selected at the same time, the estimated percentages of the times the true model is selected apart from one variable, the percentages of the times the TP are the top ranked variables among the selected, the percentage of times the TP are the top ranked variables apart from one TP, the average percentages of selected TP among the TP (sensitivity, SEN), the average percentages of selected false positive predictors (FP) among FP (false positive rate, FPR), and the average percentages of selected TP among the selected variables (positive predictive value, PPV). The mean percentages are shown along with their standard deviation (SD)

| | **LASSO** | | | | **MICE-LASSO** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Variable selection** | **Complete data** | | | | **MCAR** | | | | **MAR** | | | |
| | Best | 1 SE tolerance | 3% tolerance | 15% tolerance | Best | 1 SE tolerance | 3% tolerance | 15% tolerance | Best | 1 SE tolerance | 3% tolerance | 15% tolerance |
| | | | | | | **250 observations** | | | | | | |
| **% true models** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **% true models but 1** | 0 | 0 | 1.3 | 2.7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **% TP in top 10** | 0 | 0 | 1.0 | 2.7 | 0 | 0 | 0 | 0 | 0.3 | 0 | 0 | 0 |
| **% TP in top 10 but 1** | 17.3 | 20.3 | 20.7 | 13.7 | 4.7 | 6.3 | 4.3 | 4.3 | 6.3 | 8.0 | 7.0 | 0.3 |
| **SEN (SD)** | 92.3 (3.3) | 91.7 (3.5) | 90.5 (4.2) | 84.2 (6.0) | 95.5 (3.3) | 94.4 (3.2) | 92.8 (3.4) | 86.3 (6.9) | 96.0 (3.5) | 94.8 (3.5) | 93.3 (3.3) | 86.8 (5.6) |
| **FPR (SD)** | 35.9 (11.9) | 25.6 (10.2) | 16.6 (7.9) | 5.9 (4.6) | 86.6 (8.8) | 72.3 (12.1) | 52.6 (12.8) | 20.2 (9.2) | 85.6 (9.0) | 71.4 (12.2) | 50.7 (12.2) | 19.7 (8.8) |
| **PPV (SD)** | 62.6 (7.9) | 70.2 (8.6) | 78.2 (8.3) | 90.4 (6.7) | 40.9 (2.7) | 45.3 (4.5) | 53.1 (6.2) | 74.0 (8.9) | 41.3 (2.8) | 45.7 (4.5) | 54.2 (6.3) | 74.4 (8.2) |
| | | | | | | **1000 observations** | | | | | | |
| **% true models** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **% true models but 1** | 0 | 0.3 | 13.7 | 15.7 | 0 | 0 | 0 | 0.7 | 0 | 0 | 0 | 0 |
| **% TP in top 10** | 1.3 | 1.7 | 13.7 | 15.7 | 1.0 | 0.7 | 0 | 0.7 | 4.0 | 3.0 | 0.3 | 0 |
| **% TP in top 10 but 1** | 89.3 | 90.3 | 76.3 | 46.7 | 65.3 | 72.3 | 73.3 | 17.0 | 65.7 | 68.7 | 65.0 | 14.3 |
| **SEN (SD)** | 93.7 (1.8) | 93.5 (1.3) | 93.2 (1.3) | 88.6 (3.2) | 95.9 (3.2) | 94.9 (2.9) | 93.5 (1.3) | 89.3 (3.3) | 97.4 (3.2) | 96.5 (3.3) | 94.1 (2.2) | 89.4 (3.3) |
| **FPR (SD)** | 35.7 (11.1) | 22.1 (9.1) | 7.5 (5.0) | 2.6 (3.2) | 86.3 (7.3) | 66.7 (10.0) | 31.1 (8.3) | 10.8 (5.2) | 87.3 (7.4) | 68.5 (10.0) | 32.3 (8.3) | 12.0 (5.2) |
| **PPV (SD)** | 63.0 (7.5) | 73.5 (8.2) | 89.1 (6.6) | 95.7 (5.0) | 41.1 (2.2) | 47.4 (3.9) | 65.8 (6.2) | 84.3 (6.5) | 41.2 (2.3) | 47.1 (3.8) | 65.1 (5.9) | 82.8 (6.3) |

Table 2.8: **Variable selection** simulation study results for scenarios **S3** (assumption of moderation, without missing data) and **S4** (assumption of moderation, with missing data) for **MICE-Elasticnet** best and tolerance models in the case of **20 covariates** and 300 samples of 250 and 1000 observations. The following results are shown: the estimated percentages of the times all the 10 true predictors (TP) are selected at the same time, the estimated percentages of the times the true model is selected apart from one variable, the percentages of the times the TP are the top ranked variables among the selected, the percentage of times the TP are the top ranked variables apart from one TP, the average percentages of selected TP among the TP (sensitivity, SEN), the average percentages of selected false positive predictors (FP) among FP (false positive rate, FPR), and the average percentages of selected TP among the selected variables (positive predictive value, PPV). The mean percentages are shown along with their standard deviation (SD)

| | **ELASTICNET** | | | | **MICE-ELASTICNET** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Variable selection** | **Complete data** | | | | **MCAR** | | | | **MAR** | | | |
| | Best | 1 SE tolerance | 3% tolerance | 15% tolerance | Best | 1 SE tolerance | 3% tolerance | 15% tolerance | Best | 1 SE tolerance | 3% tolerance | 15% tolerance |
| | | | | | | **250 observations** | | | | | | |
| **% true models** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **% true models but 1** | 0 | 1.3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **% TP in top 10** | 0 | 1.3 | 0 | 0 | 0 | 0 | 0 | 0 | 0.3 | 0 | 0 | 0 |
| **% TP in top 10 but 1** | 15.3 | 12.3 | 2.3 | 0.3 | 4.3 | 3.3 | 1.3 | 0 | 5.3 | 5.0 | 0.7 | 0 |
| **SEN (SD)** | 92.7 (3.3) | 89.3 (5.1) | 76.8 (14.8) | 41.9 (27.4) | 96.7 (3.4) | 94.7 (4.1) | 92.8 (5.2) | 84.3 (13.6) | 96.9 (3.5) | 95.4 (4.1) | 93.4 (5.5) | 84.8 (15.9) |
| **FPR (SD)** | 40.9 (14.9) | 18.2 (16.7) | 11.2 (15.0) | 5.1 (8.9) | 92.8 (7.4) | 74.7 (20.1) | 59.6 (23.9) | 34.0 (18.9) | 92.5 (7.5) | 74.5 (20.7) | 60.8 (24.1) | 35.0 (19.3) |
| **PPV** (SD) | 59.9 (8.7) | 78.6 (13.6) | 85.8 (14.0) | 90.6 (11.5) | 39.5 (2.1) | 45.4 (7.5) | 51.8 (11.6) | 64.4 (13.0) | 39.6 (2.1) | 45.7 (8.1) | 51.6 (12.0) | 64.0 (13.1) |
| | | | | | | **1000 observations** | | | | | | |
| **% true models** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **% true models but 1** | 0 | 8.3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.3 | 0 |
| **% TP in top 10** | 1.3 | 8.3 | 0 | 0 | 1.0 | 0 | 0 | 0 | 4.0 | 0.3 | 0.3 | 0 |
| **% TP in top 10 but 1** | 89.0 | 80.0 | 25.0 | 0 | 65.7 | 64.7 | 13.0 | 0 | 65.3 | 56.3 | 8.7 | 0 |
| **SEN (SD)** | 93.7 (1.8) | 93.3 (1.0) | 81.6 (5.0) | 20.3 (5.2) | 95.9 (3.2) | 93.6 (1.4) | 84.9 (5.6) | 32.8 (21.1) | 97.5 (3.2) | 94.0 (2.4) | 86.5 (5.2) | 36.7 (23.2) |
| **FPR (SD)** | 36.3 (11.5) | 8.4 (4.9) | 1.5 (2.3) | 0.1 (0.7) | 87.1 (7.5) | 30.8 (13.0) | 8.6 (10.3) | 2.4 (5.4) | 87.9 (7.4) | 32.7 (14.6) | 10.6 (11.0) | 3.2 (6.5) |
| **PPV (SD)** | 62.7 (7.5) | 87.9 (6.3) | 97.3 (4.1) | 99.3 (4.0) | 40.9 (2.3) | 66.7 (8.1) | 88.0 (10.3) | 94.9 (8.8) | 41.0 (2.3) | 65.6 (8.4) | 85.6 (10.6) | 93.6 (9.5) |

Table 2.9: **Accuracy** simulation study results **MICE-Lasso** analysis with Harrell bootstrap validation: scenarios **S3** (assumption of moderation, without missing data) and **S4** (assumption of moderation, with missing data, complete outcome) based on 300 data sets of **20 variables** each (**n=250**). Means of all estimates along with their corresponding 2.5th and 97.5th percentile values within parenthesis. The theoretical MSE is $1.74^2 = 3.028$

| **Estimates** | **Complete cases** | | | |
|---|---|---|---|---|
| | Best model | 1 SE tolerance | 3% tolerance | 15% tolerance |
| $MSE_{apparent}$ | 2.764 (2.232,3.313) | 2.873 (2.328,3.483) | 3.076 (2.485,3.729) | 3.939 (3.121,4.889) |
| $\beta_{LP}$ | 1.062 (1.044,1.079) | 1.089 (1.069,1.112) | 1.132 (1.102,1.165) | 1.259 (1.204,1.332) |
| Tuning $\lambda$ | 0.077 (0.054,0.095) | 0.111 (0.085,0.134) | 0.165 (0.134,0.209) | 0.328 (0.262,0.410) |
| $MSE_{ext}$ | 3.482 (3.266,3.794) | 3.557 (3.313,3.914) | 3.759 (3.436,4.214) | 4.799 (4.070,5.835) |
| $Optimism_{ext}$ | -0.719 (-1.300,-0.093) | -0.684 (-1.283,-0.050) | -0.683 (-1.281,0.017) | -0.860 (-1.706,-0.064) |
| $Optimism_{int}$ | -0.646 (-0.806,-0.521) | -0.578 (-0.731,-0.462) | -0.523 (-0.665,-0.420) | -0.486 (-0.620,-0.373) |
| $MSE_{corrected}$ | 3.410 (2.788,4.131) | 3.451 (2.826,4.194) | 3.598 (2.937,4.384) | 4.425 (3.541,5.495) |
| $\beta_{LP^*}$ | 1.025 (1.012,1.036) | 1.052 (1.035,1.069) | 1.092 (1.067,1.118) | 1.216 (1.167,1.273) |
| | **MCAR** | | | |
| $MSE_{apparent}$ | 3.105 (2.347,3.918) | 3.214 (2.413,4.079) | 3.435 (2.570,4.381) | 4.444 (3.260,5.729) |
| $\beta_{LP}$ | 1.054 (1.034,1.076) | 1.083 (1.057,1.111) | 1.128 (1.093,1.169) | 1.274 (1.198,1.371) |
| Tuning $\lambda$ | 0.072 (0.049,0.099) | 0.107 (0.077,0.145) | 0.165 (0.123,0.221) | 0.358 (0.270,0.475) |
| $MSE_{ext}$ | 3.800 (3.438,4.254) | 3.876 (3.473,4.419) | 4.109 (3.592,4.898) | 5.359 (4.241,6.776) |
| $Optimism_{ext}$ | -0.695 (-1.513,0.125) | -0.662 (-1.457,0.174) | -0.674 (-1.502,0.209) | -0.914 (-1.944,0.197) |
| $Optimism_{int}$ | -1.014 (-1.356,-0.770) | -0.908 (-1.210,-0.688) | -0.798 (-1.065,-0.608) | -0.636 (-0.857,-0.475) |
| $MSE_{corrected}$ | 4.119 (3.168,5.157) | 4.122 (3.150,5.159) | 4.234 (3.208,5.322) | 5.080 (3.784,6.407) |
| $\beta_{LP^*}$ | 0.992 (0.965,1.016) | 1.022 (0.993,1.047) | 1.067 (1.032,1.096) | 1.204 (1.148,1.272) |
| | **MAR** | | | |
| $MSE_{apparent}$ | 3.078 (2.460,3.814) | 3.188 (2.556,3.947) | 3.412 (2.720,4.289) | 4.432 (3.452,5.587) |
| $\beta_{LP}$ | 1.055 (1.038,1.075) | 1.084 (1.061,1.109) | 1.129 (1.098,1.168) | 1.272 (1.207,1.361) |
| Tuning $\lambda$ | 0.071 (0.051,0.098) | 0.106 (0.077,0.140) | 0.163 (0.121,0.214) | 0.353 (0.266,0.471) |
| $MSE_{ext}$ | 4.054 (3.620,4.593) | 4.162 (3.682,4.745) | 4.426 (3.862,5.147) | 5.705 (4.695,6.957) |
| $Optimism_{ext}$ | -0.976 (-1.763,-0.196) | -0.973 (-1.738,-0.212) | -1.014 (-1.770,-0.274) | -1.273 (-2.279,-0.292) |
| $Optimism_{int}$ | -0.909 (-1.162,-0.683) | -0.823 (-1.056,-0.615) | -0.737 (-0.959,-0.543) | -0.615 (-0.805,-0.453) |
| $MSE_{corrected}$ | 3.986 (3.190,4.882) | 4.011 (3.206,4.908) | 4.149 (3.304,5.089) | 5.046 (3.959,6.233) |
| $\beta_{LP^*}$ | 1.004 (0.986,1.022) | 1.034 (1.013,1.057) | 1.077 (1.051,1.108) | 1.213 (1.166,1.278) |

Table 2.10: **Accuracy** simulation study results for **MICE-Elasticnet** analysis with Harrell bootstrap validation: scenarios **S3** (assumption of moderation, without missing data) and **S4** (assumption of moderation, with missing data, complete outcome) based on 300 data sets of **20 variables** each (**n=250**). Means of all estimates along with their corresponding 2.5th and 97.5th percentile values within parenthesis. The theoretical MSE is $1.74^2 = 3.028$

| **Estimates** | **Complete cases** | | | |
|---|---|---|---|---|
| | Best model | 1 SE tolerance | 3% tolerance | 15% tolerance |
| $MSE_{apparent}$ | 2.761 (2.220,3.287) | 3.411 (2.492,4.631) | 4.858 (2.796,7.019) | 8.191 (3.771,11.757) |
| $\beta_{LP}$ | 1.068 (1.050,1.088) | 1.197 (1.100,1.326) | 1.370 (1.147,1.600) | 2.059 (1.351,3.554) |
| Tuning $\alpha$ | 0.581 (0.100,0.900) | 0.899 (0.900,0.900) | 0.900 (0.900,0.900) | 0.900 (0.900,0.900) |
| Tuning $\lambda$ | 0.149 (0.083,0.357) | 0.398 (0.220,0.580) | 0.738 (0.580,0.941) | 1.708 (1.199,1.947) |
| $MSE_{ext}$ | 3.495 (3.260,3.792) | 4.178 (3.500,5.471) | 5.889 (3.716,8.452) | 9.596 (4.819,13.304) |
| $Optimism_{ext}$ | -0.734 (-1.343,-0.096) | -0.768 (-1.475,-0.044) | -1.032 (-2.216,0.032) | -1.405 (-3.321,0.230) |
| $Optimism_{int}$ | -0.653 (-0.817,-0.534) | -0.564 (-0.721,-0.447) | -0.519 (-0.654,-0.408) | -0.430 (-0.563,-0.278) |
| $MSE_{corrected}$ | 3.414 (2.752,4.096) | 3.975 (2.998,5.333) | 5.376 (3.310,7.548) | 8.621 (4.253,12.151) |
| $\beta_{LP^*}$ | 1.035 (1.022,1.047) | 1.109 (1.069,1.154) | 1.200 (1.123,1.294) | 1.534 (1.304,1.863) |
| | **MCAR** | | | |
| $MSE_{apparent}$ | 3.104 (2.351,3.916) | 3.537 (2.685,4.651) | 4.383 (3.099,6.144) | 6.964 (4.354,10.529) |
| $\beta LP$ | 1.064 (1.039,1.090) | 1.170 (1.104,1.256) | 1.312 (1.170,1.516) | 1.861 (1.390,2.964) |
| Tuning $\alpha$ | 0.413 (0.120,0.810) | 0.877 (0.735,0.900) | 0.900 (0.900,0.900) | 0.900 (0.900,0.900) |
| Tuning $\lambda$ | 0.203 (0.081,0.387) | 0.448 (0.257,0.675) | 0.780 (0.534,1.074) | 1.873 (1.301,2.496) |
| $MSE_{ext}$ | 3.810 (3.446,4.243) | 4.255 (3.630,5.132) | 5.260 (3.968,7.149) | 8.214 (5.368,11.979) |
| $Optimism_{ext}$ | -0.705 (-1.514,0.134) | -0.717 (-1.493,0.198) | -0.876 (-1.923,0.179) | -1.250 (-2.783,0.033) |
| $Optimism_{int}$ | -1.004 (-0.817,-0.534) | -0.819 (-0.721,-0.447) | -0.695 (-0.654,-0.408) | -0.523 (-0.563,-0.278) |
| $MSE_{corrected}$ | 4.109 (3.172,5.145) | 4.356 (3.327,5.644) | 5.078 (3.727,6.846) | 7.487 (4.925,11.087) |
| $\beta_{LP^*}$ | 1.003 (1.022,1.047) | 1.080 (1.069,1.154) | 1.177 (1.123,1.294) | 1.526 (1.304,1.863) |
| | **MAR** | | | |
| $MSE_{apparent}$ | 3.078 (2.454,3.816) | 3.492 (2.710,4.423) | 4.271 (3.076,6.169) | 6.690 (4.262,10.325) |
| $\beta_{LP}$ | 1.066 (1.047,1.089) | 1.167 (1.099,1.243) | 1.299 (1.156,1.477) | 1.788 (1.382,2.605) |
| Tuning $\alpha$ | 0.378 (0.115,0.770) | 0.874 (0.690,0.900) | 0.900 (0.900,0.900) | 0.900 (0.900,0.900) |
| Tuning $\lambda$ | 0.210 (0.086,0.370) | 0.453 (0.267,0.639) | 0.781 (0.531,1.035) | 1.856 (1.364,2.443) |
| $MSE_{ext}$ | 3.774 (3.444,4.229) | 4.193 (3.662,5.042) | 5.120 (3.880,7.216) | 7.911 (5.031,11.657) |
| $Optimism_{ext}$ | -0.696 (-1.504,0.067) | -0.702 (-1.494,0.040) | -0.849 (-1.833,0.023) | -1.221 (-2.764,0.090) |
| $Optimism_{int}$ | -0.902 (-1.160,-0.673) | -0.761 (-0.975,-0.555) | -0.667 (-0.861,-0.486) | -0.522 (-0.713,-0.356) |
| $MSE_{corrected}$ | 3.980 (3.170,4.869) | 4.253 (3.336,5.277) | 4.938 (3.682,6.816) | 7.212 (4.804,10.838) |
| $\beta_{LP^*}$ | 1.016 (0.995,1.036) | 1.090 (1.055,1.132) | 1.183 (1.121,1.274) | 1.522 (1.324,1.872) |

Table 2.11: **Variable selection** simulation study results for scenarios **S3** (assumption of moderation, without missing data) and **S5** (assumption of moderation, missing data also in the outcome) for **MICE-Lasso** best and tolerance models in the case of **20 covariates** and 300 samples of 250 and 1000 observations. The following results are shown: the estimated percentages of the times all the 10 true predictors (TP) are selected at the same time, the estimated percentages of the times the true model is selected apart from one variable, the percentages of the times the TP are the top ranked variables among the selected, the percentage of times the TP are the top ranked variables apart from one TP, the average percentages of selected TP among the TP (sensitivity, SEN), the average percentages of selected false positive predictors (FP) among FP (false positive rate, FPR), and the average percentages of selected TP among the selected variables (positive predictive value, PPV). The mean percentages are shown along with their standard deviation (SD)

| | **LASSO** | | | | **MICE-LASSO** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Variable selection** | **Complete data** | | | | **MCAR** | | | | **MAR** | | | |
| | Best | 1 SE tolerance | 3% tolerance | 15% tolerance | Best | 1 SE tolerance | 3% tolerance | 15% tolerance | Best | 1 SE tolerance | 3% tolerance | 15% tolerance |
| | | | | | | **250 observations** | | | | | | |
| **% true models** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **% true models but 1** | 0 | 0 | 1.3 | 2.7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **% TP in top 10** | 0 | 0 | 1.0 | 2.7 | 0 | 0 | 0 | 0 | 0.3 | 0.3 | 0 | 0 |
| **% TP in top 10 but 1** | 17.3 | 20.3 | 20.7 | 13.7 | 0 | 0.7 | 1.7 | 0.7 | 3.0 | 4.7 | 4.3 | 0.7 |
| **SEN (SD)** | 92.3 (3.3) | 91.7 (3.5) | 90.5 (4.2) | 84.2 (6.0) | 96.6 (3.4) | 95.1 (3.5) | 93.7 (3.4) | 87.5 (6.4) | 97.5 (3.2) | 96.6 (3.7) | 95.2 (4.1) | 88.9 (6.1) |
| **FPR (SD)** | 35.9 (11.9) | 25.6 (10.2) | 16.6 (7.9) | 5.9 (4.6) | 94.1 (5.0) | 84.4 (8.5) | 67.4 (11.8) | 30.2 (11.5) | 93.6 (5.3) | 84.3 (8.8) | 67.8 (12.2) | 31.0 (11.6) |
| **PPV (SD)** | 62.6 (7.9) | 70.2 (8.6) | 78.2 (8.3) | 90.4 (6.7) | 39.1 (1.5) | 41.5 (2.7) | 46.9 (4.6) | 65.6 (8.2) | 39.5 (1.7) | 41.9 (2.8) | 47.2 (4.9) | 65.4 (8.3) |
| | | | | | | **1000 observations** | | | | | | |
| **% true models** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **% true models but 1** | 0 | 0.3 | 13.7 | 15.7 | 0 | 0 | 0 | 0.3 | 0 | 0 | 0 | 0 |
| **% TP in top 10** | 1.3 | 1.7 | 13.7 | 15.7 | 0.7 | 0.3 | 0 | 0.3 | 11.3 | 9.7 | 1.0 | 0 |
| **% TP in top 10 but 1** | 89.3 | 90.3 | 76.3 | 46.7 | 37.3 | 43.7 | 43.0 | 8.7 | 47.3 | 46.3 | 37.0 | 3.3 |
| **SEN (SD)** | 93.7 (1.8) | 93.5 (1.3) | 93.2 (1.3) | 88.6 (3.2) | 97.0 (3.3) | 96.0 (3.3) | 93.9 (1.9) | 89.6 (3.6) | 99.2 (2.1) | 99.0 (2.4) | 96.8 (3.3) | 90.4 (3.5) |
| **FPR (SD)** | 35.7 (11.1) | 22.1 (9.1) | 7.5 (5.0) | 2.6 (3.2) | 93.6 (4.8) | 79.6 (7.4) | 42.5 (9.8) | 13.9 (5.4) | 94.9 (5.0) | 82.5 (7.6) | 47.0 (9.6) | 17.6 (5.2) |
| **PPV (SD)** | 63.0 (7.5) | 73.5 (8.2) | 89.1 (6.6) | 95.7 (5.0) | 39.4 (1.5) | 43.1 (2.5) | 58.6 (5.6) | 80.6 (6.2) | 39.6 (1.4) | 43.0 (2.4) | 56.7 (5.1) | 76.7 (5.2) |

Table 2.12: **Variable selection** simulation study results for scenarios **S3** (assumption of moderation, without missing data) and **S5** (assumption of moderation, missing data also in the outcome) for **Elasticnet** and **MICE-Elasticnet** best and tolerance models in the case of **20 covariates** and 300 samples of 250 and 1000 observations. The following results are shown: the estimated percentages of the times all the 10 true predictors (TP) are selected at the same time, the estimated percentages of the times the true model is selected apart from one variable, the percentages of the times the TP are the top ranked variables among the selected, the percentage of times the TP are the top ranked variables apart from one TP, the average percentages of selected TP among the TP (sensitivity, SEN), the average percentages of selected false positive predictors (FP) among FP (false positive rate, FPR), and the average percentages of selected TP among the selected variables (positive predictive value, PPV). The mean percentages are shown along with their standard deviation (SD)

| | **ELASTICNET** | | | | **MICE-ELASTICNET** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Variable selection** | **Complete data** | | | | **MCAR** | | | | **MAR** | | | |
| | Best | 1 SE tolerance | 3% tolerance | 15% tolerance | Best | 1 SE tolerance | 3% tolerance | 15% tolerance | Best | 1 SE tolerance | 3% tolerance | 15% tolerance |
| | | | | | | **250 observations** | | | | | | |
| **% true models** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **% true models but 1** | 0 | 1.3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **% TP in top 10** | 0 | 1.3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **% TP in top 10 but 1** | 15.3 | 12.3 | 2.3 | 0.3 | 0.3 | 1.0 | 1.0 | 0 | 3.7 | 3.3 | 1.7 | 0 |
| **SEN (SD)** | 92.7 (3.3) | 89.3 (5.1) | 76.8 (14.8) | 41.9 (27.4) | 97.8 (3.2) | 95.8 (3.7) | 94.0 (4.7) | 87.0 (11.4) | 98.5 (2.9) | 97.2 (3.9) | 95.8 (4.8) | 89.6 (11.9) |
| **FPR (SD)** | 40.9 (14.9) | 18.2 (16.7) | 11.2 (15.0) | 5.1 (8.9) | 97.5 (3.6) | 86.0 (13.6) | 72.7 (20.1) | 43.2 (20.0) | 97.5 (3.6) | 87.0 (14.3) | 74.8 (21.2) | 47.6 (21.0) |
| **PPV (SD)** | 59.9 (8.7) | 78.6 (13.6) | 85.8 (14.0) | 90.6 (11.5) | 38.5 (1.1) | 41.4 (4.2) | 45.9 (8.0) | 58.7 (12.1) | 38.7 (1.1) | 41.6 (4.6) | 45.9 (8.6) | 57.0 (11.8) |
| | | | | | | **1000 observations** | | | | | | |
| **% true models** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **% true models but 1** | 0 | 8.3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **% TP in top 10** | 1.3 | 8.3 | 0 | 0 | 0.3 | 0 | 0 | 0.0 | 11.3 | 0.7 | 0.0 | 0.0 |
| **% TP in top 10 but 1** | 89.0 | 80.0 | 25.0 | 0.0 | 38.3 | 37.3 | 4.7 | 0.0 | 47.3 | 28.0 | 1.0 | 0.0 |
| **SEN (SD)** | 93.7 (1.8) | 93.3 (1.0) | 81.6 (5.0) | 20.3 (5.2) | 97.1 (3.3) | 94.0 (2.2) | 86.4 (6.5) | 42.5 (26.1) | 99.3 (2.0) | 97.0 (3.3) | 90.0 (5.7) | 57.4 (27.3) |
| **FPR (SD)** | 36.3 (11.5) | 8.4 (4.9) | 1.5 (2.3) | 0.1 (0.7) | 94.1 (5.0) | 43.0 (15.9) | 13.9 (12.6) | 4.4 (7.2) | 95.6 (4.7) | 51.1 (18.7) | 23.8 (16.9) | 9.4 (10.3) |
| **PPV (SD)** | 62.7 (7.5) | 87.9 (6.3) | 97.3 (4.1) | 99.3 (4.0) | 39.2 (1.5) | 59.0 (8.3) | 81.9 (11.4) | 91.6 (10.2) | 39.4 (1.3) | 55.7 (8.7) | 73.1 (12.0) | 85.1 (11.5) |

Table 2.13: **Accuracy** simulation study results for **MICE-Lasso** analysis with Harrell bootstrap validation: scenarios **S3** (assumption of moderation, complete data) and **S5** (assumption of moderation, missing data also in the outcome) based on 300 data sets of **20 variables** each (**n=250**). Means of all estimates along with their corresponding 2.5th and 97.5th percentile values within parenthesis. The theoretical MSE is $1.74^2 = 3.028$

| **Estimates** | **Complete cases** | | | |
|---|---|---|---|---|
| | Best model | 1 SE tolerance | 3% tolerance | 15% tolerance |
| $MSE_{apparent}$ | 2.764 (2.232,3.313) | 2.873 (2.328,3.483) | 3.076 (2.485,3.729) | 3.939 (3.121,4.889) |
| $\beta_{LP}$ | 1.062 (1.044,1.079) | 1.089 (1.069,1.112) | 1.132 (1.102,1.165) | 1.259 (1.204,1.332) |
| Tuning $\lambda$ | 0.077 (0.054,0.095) | 0.111 (0.085,0.134) | 0.165 (0.134,0.209) | 0.328 (0.262,0.410) |
| $MSE_{ext}$ | 3.482 (3.266,3.794) | 3.557 (3.313,3.914) | 3.759 (3.436,4.214) | 4.799 (4.070,5.835) |
| $Optimism_{ext}$ | -0.719 (-1.300,-0.093) | -0.684 (-1.283,-0.050) | -0.683 (-1.281,0.017) | -0.860 (-1.706,-0.064) |
| $Optimism_{int}$ | -0.646 (-0.806,-0.521) | -0.578 (-0.731,-0.462) | -0.523 (-0.665,-0.420) | -0.486 (-0.620,-0.373) |
| $MSE_{corrected}$ | 3.410 (2.788,4.131) | 3.451 (2.826,4.194) | 3.598 (2.937,4.384) | 4.425 (3.541,5.495) |
| $\beta_{LP^*}$ | 1.025 (1.012,1.036) | 1.052 (1.035,1.069) | 1.092 (1.067,1.118) | 1.216 (1.167,1.273) |
| | **MCAR** | | | |
| $MSE_{apparent}$ | 3.312 (2.424,4.240) | 3.415 (2.488,4.376) | 3.633 (2.614,4.699) | 4.669 (3.244,6.144) |
| $\beta_{LP}$ | 1.053 (1.031,1.075) | 1.083 (1.053,1.112) | 1.130 (1.085,1.176) | 1.283 (1.195,1.384) |
| Tuning $\lambda$ | 0.070 (0.047,0.096) | 0.105 (0.072,0.141) | 0.163 (0.114,0.218) | 0.362 (0.249,0.490) |
| $MSE_{ext}$ | 4.036 (3.606,4.610) | 4.092 (3.642,4.701) | 4.301 (3.734,5.110) | 5.534 (4.362,6.987) |
| $Optimism_{ext}$ | -0.724 (-1.798,0.310) | -0.678 (-1.742,0.314) | -0.669 (-1.692,0.316) | -0.866 (-2.007,0.265) |
| $Optimism_{int}$ | -1.264 (-1.709,-0.903) | -1.150 (-1.551,-0.807) | -1.022 (-1.390,-0.707) | -0.794 (-1.067,-0.537) |
| $MSE_{corrected}$ | 4.576 (3.387,5.892) | 4.565 (3.371,5.881) | 4.655 (3.411,5.994) | 5.463 (3.893,7.044) |
| $\beta_{LP^*}$ | 0.980 (0.946,1.013) | 1.010 (0.978,1.044) | 1.055 (1.017,1.094) | 1.196 (1.129,1.271) |
| | **MAR** | | | |
| $MSE_{apparent}$ | 3.286 (2.503,4.187) | 3.392 (2.585,4.314) | 3.615 (2.756,4.610) | 4.675 (3.542,6.013) |
| $\beta_{LP}$ | 1.055 (1.037,1.075) | 1.084 (1.059,1.111) | 1.130 (1.095,1.171) | 1.280 (1.209,1.382) |
| Tuning $\lambda$ | 0.069 (0.048,0.094) | 0.104 (0.074,0.139) | 0.161 (0.119,0.214) | 0.356 (0.267,0.484) |
| $MSE_{ext}$ | 3.977 (3.536,4.609) | 4.029 (3.599,4.667) | 4.226 (3.695,5.021) | 5.437 (4.378,6.791) |
| $Optimism_{ext}$ | -0.692 (-1.727,0.303) | -0.637 (-1.643,0.333) | -0.611 (-1.651,0.359) | -0.762 (-1.964,0.375) |
| $Optimism_{int}$ | -1.099 (-1.449,-0.774) | -1.012 (-1.338,-0.705) | -0.917 (-1.213,-0.624) | -0.756 (-1.002,-0.496) |
| $MSE_{corrected}$ | 4.385 (3.320,5.644) | 4.404 (3.330,5.651) | 4.532 (3.427,5.809) | 5.431 (4.087,7.028) |
| $\beta_{LP^*}$ | 0.997 (0.968,1.022) | 1.026 (0.999,1.053) | 1.069 (1.040,1.103) | 1.208 (1.157,1.278) |

Table 2.14: **Accuracy** simulation study results for **MICE-Elasticnet** analysis with Harrell bootstrap validation: scenarios **S3** (assumption of moderation, complete data) and **S5** (assumption of moderation, missing data also in the outcome ), based on 300 data sets of **20 variables** each (**n=250**). Means of all estimates along with their corresponding 2.5th and 97.5th percentile values within parenthesis. The theoretical MSE is $1.74^2 = 3.028$

| **Estimates** | **Complete cases** | | | |
|---|---|---|---|---|
| | Best model | 1 SE tolerance | 3% tolerance | 15% tolerance |
| $\text{MSE}_{apparent}$ | 2.761 (2.220,3.287) | 3.411 (2.492,4.631) | 4.858 (2.796,7.019) | 8.191 (3.771,11.757) |
| $\beta_{LP}$ | 1.068 (1.050,1.088) | 1.197 (1.100,1.326) | 1.370 (1.147,1.600) | 2.059 (1.351,3.554) |
| Tuning $\alpha$ | 0.581 (0.100,0.900) | 0.899 (0.900,0.900) | 0.900 (0.900,0.900) | 0.900 (0.900,0.900) |
| Tuning $\lambda$ | 0.149 (0.083,0.357) | 0.398 (0.220,0.580) | 0.738 (0.580,0.941) | 1.708 (1.199,1.947) |
| $\text{MSE}_{ext}$ | 3.495 (3.260,3.792) | 4.178 (3.500,5.471) | 5.889 (3.716,8.452) | 9.596 (4.819,13.304) |
| $\text{Optimism}_{ext}$ | -0.734 (-1.343,-0.096) | -0.768 (-1.475,-0.044) | -1.032 (-2.216,0.032) | -1.405 (-3.321,0.230) |
| $\text{Optimism}_{int}$ | -0.653 (-0.817,-0.534) | -0.564 (-0.721,-0.447) | -0.519 (-0.654,-0.408) | -0.430 (-0.563,-0.278) |
| $\text{MSE}_{corrected}$ | 3.414 (2.752,4.096) | 3.975 (2.998,5.333) | 5.376 (3.310,7.548) | 8.621 (4.253,12.151) |
| $\beta_{LP^*}$ | 1.035 (1.022,1.047) | 1.109 (1.069,1.154) | 1.200 (1.123,1.294) | 1.534 (1.304,1.863) |
| | **MCAR** | | | |
| $\text{MSE}_{apparent}$ | 3.314 (2.421,4.251) | 3.722 (2.728,4.761) | 4.514 (3.228,6.048) | 7.071 (4.670,9.899) |
| $\beta LP$ | 1.064 (1.035,1.092) | 1.170 (1.104,1.255) | 1.313 (1.174,1.490) | 1.861 (1.406,2.594) |
| Tuning $\alpha$ | 0.415 (0.130,0.780) | 0.871 (0.730,0.900) | 0.900 (0.900,0.900) | 0.900 (0.900,0.900) |
| Tuning $\lambda$ | 0.202 (0.078,0.379) | 0.440 (0.225,0.665) | 0.767 (0.469,1.070) | 1.872 (1.228,2.618) |
| $\text{MSE}_{ext}$ | 4.032 (3.598,4.597) | 4.402 (3.769,5.257) | 5.316 (4.104,6.967) | 8.260 (5.471,11.412) |
| $\text{Optimism}_{ext}$ | -0.718 (-1.793,0.317) | -0.680 (-1.667,0.346) | -0.802 (-1.911,0.397) | -1.190 (-2.846,0.297) |
| $\text{Optimism}_{int}$ | -1.249 (-1.682,-0.895) | -1.016 (-1.373,-0.707) | -0.845 (-1.136,-0.570) | -0.604 (-0.869,-0.374) |
| $\text{MSE}_{corrected}$ | 4.563 (3.500,5.672) | 4.738 (3.607,5.927) | 5.359 (3.978,6.912) | 7.674 (5.238,10.469) |
| $\beta_{LP^*}$ | 0.989 (0.954,1.024) | 1.073 (1.032,1.113) | 1.180 (1.115,1.254) | 1.561 (1.363,1.843) |
| | **MAR** | | | |
| $\text{MSE}_{apparent}$ | 3.292 (2.507,4.206) | 3.666 (2.774,4.680) | 4.363 (3.157,5.856) | 6.681 (4.373,9.718) |
| $\beta_{LP}$ | 1.067 (1.045,1.091) | 1.162 (1.098,1.242) | 1.288 (1.162,1.459) | 1.752 (1.372,2.553) |
| Tuning $\alpha$ | 0.355 (0.110,0.745) | 0.857 (0.655,0.900) | 0.900 (0.900,0.900) | 0.900 (0.900,0.900) |
| Tuning $\lambda$ | 0.221 (0.091,0.403) | 0.456 (0.279,0.670) | 0.775 (0.539,1.070) | 1.862 (1.350,2.564) |
| $\text{MSE}_{ext}$ | 3.969 (3.531,4.604) | 4.270 (3.704,5.229) | 5.032 (3.988,6.787) | 7.624 (5.266,10.951) |
| $\text{Optimism}_{ext}$ | -0.678 (-1.673,0.313) | -0.605 (-1.693,0.361) | -0.669 (-1.828,0.361) | -0.943 (-2.477,0.530) |
| $\text{Optimism}_{int}$ | -0.831 (-1.431,-0.771) | -0.737 (-1.218,-0.630) | -0.659 (-1.067,-0.534) | -0.547 (-0.888,-0.348) |
| $\text{MSE}_{corrected}$ | 4.400 (3.310,5.627) | 4.639 (3.487,5.882) | 5.311 (3.838,6.693) | 7.681 (4.934,10.465) |
| $\beta_{LP^*}$ | 1.007 (0.979,1.034) | 1.086 (1.052,1.127) | 1.186 (1.125,1.267) | 1.541 (1.369,1.807) |

Table 2.15: **Variable selection** simulation study results for scenarios **S3** (assumption of moderation, complete data) and **S6** (assumption of moderation, missing data, complete outcome, interaction terms in imputation model) for **MICE-Lasso** best and tolerance models in the case of **20 covariates** and 300 samples of 250 and 1000 observations. The following results are shown: the estimated percentages of the times all the 10 true predictors (TP) are selected at the same time, the estimated percentages of the times the true model is selected apart from one variable, the percentages of the times the TP are the top ranked variables among the selected, the percentage of times the TP are the top ranked variables apart from one TP, the average percentages of selected TP among the TP (sensitivity, SEN), the average percentages of selected false positive predictors (FP) among FP (false positive rate, FPR), and the average percentages of selected TP among the selected variables (positive predictive value, PPV). The mean percentages are shown along with their standard deviation (SD)

| | **LASSO** | | | | **MICE-LASSO** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Variable selection** | **Complete data** | | | | **MCAR** | | | | **MAR** | | | |
| | Best | 1 SE tolerance | 3% tolerance | 15% tolerance | Best | 1 SE tolerance | 3% tolerance | 15% tolerance | Best | 1 SE tolerance | 3% tolerance | 15% tolerance |
| | | | | | | **250 observations** | | | | | | |
| **% true models** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **% true models but 1** | 0 | 0 | 1.3 | 2.7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **% TP in top 10** | 0 | 0 | 1.0 | 2.7 | 0 | 0 | 0 | 0 | 0.3 | 0 | 0 | 0 |
| **% TP in top 10 but 1** | 17.3 | 20.3 | 20.7 | 13.7 | 2.7 | 3.0 | 3.7 | 3.7 | 2.0 | 5.0 | 5.7 | 3.0 |
| **SEN (SD)** | 92.3 (3.3) | 91.7 (3.5) | 90.5 (4.2) | 84.2 (6.0) | 98.6 (2.7) | 97.2 (3.5) | 94.8 (3.6) | 88.7 (5.8) | 98.2 (3.2) | 96.4 (3.6) | 94.2 (3.6) | 88.1 (6.5) |
| **FPR (SD)** | 35.9 (11.9) | 25.6 (10.2) | 16.6 (7.9) | 5.9 (4.6) | 98.4 (3.7) | 93.5 (9.4) | 77.5 (18.4) | 30.9 (17.2) | 96.5 (5.8) | 88.7 (11.6) | 69.2 (16.8) | 24.5 (10.9) |
| **PPV (SD)** | 62.6 (7.9) | 70.2 (8.6) | 78.2 (8.3) | 90.4 (6.7) | 38.5 (1.0) | 39.5 (2.6) | 44.2 (6.2) | 66.5 (11.1) | 38.9 (1.6) | 40.7 (3.5) | 46.8 (6.4) | 70.4 (8.9) |
| | | | | | | **1000 observations** | | | | | | |
| **% true models** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **% true models but 1** | 0 | 0.3 | 13.7 | 15.7 | 0 | 0 | 0 | 1.0 | 0 | 0 | 0 | 1.0 |
| **% TP in top 10** | 1.3 | 1.7 | 13.7 | 15.7 | 0.7 | 0.7 | 0 | 1.0 | 1.3 | 0.7 | 0 | 1.0 |
| **% TP in top 10 but 1** | 89.3 | 90.3 | 76.3 | 46.7 | 34.7 | 51.7 | 65.0 | 25.0 | 48.3 | 62.7 | 67.0 | 24.3 |
| **SEN (SD)** | 93.7 (1.8) | 93.5 (1.3) | 93.2 (1.3) | 88.6 (3.2) | 97.5 (3.2) | 95.5 (3.1) | 93.7 (1.6) | 89.9 (3.4) | 97.0 (3.3) | 95.5 (3.1) | 93.7 (1.6) | 89.5 (3.4) |
| **FPR (SD)** | 35.7 (11.1) | 22.1 (9.1) | 7.5 (5.0) | 2.6 (3.2) | 97.4 (4.1) | 85.4 (10.3) | 38.2 (11.5) | 10.7 (5.6) | 96.7 (4.0) | 83.3 (9.6) | 36.9 (10.5) | 10.9 (5.5) |
| **PPV (SD)** | 63.0 (7.5) | 73.5 (8.2) | 89.1 (6.6) | 95.7 (5.0) | 38.5 (1.3) | 41.3 (3.1) | 61.4 (7.1) | 84.6 (7.1) | 38.5 (1.2) | 41.9 (2.9) | 62.1 (6.6) | 84.3 (6.9) |

Table 2.16: **Accuracy** simulation study results for **MICE-Lasso** analysis with Harrell bootstrap validation: scenarios S3 (assumption of moderation, without missing data) and **S6** (assumption of moderation, missing data, interaction terms in the imputation model), based on 300 data sets of **20 variables** each (**n=250**). Means of all estimates along with their corresponding 2.5th and 97.5th percentile values within parenthesis. The theoretical MSE is $1.74^2 = 3.028$

| **Estimates** | **Complete cases** | | | |
|---|---|---|---|---|
| | Best model | 1 SE tolerance | 3% tolerance | 15% tolerance |
| $MSE_{apparent}$ | 2.764 (2.232,3.313) | 2.873 (2.328,3.483) | 3.076 (2.485,3.729) | 3.939 (3.121,4.889) |
| $\beta_{LP}$ | 1.062 (1.044,1.079) | 1.089 (1.069,1.112) | 1.132 (1.102,1.165) | 1.259 (1.204,1.332) |
| Tuning $\lambda$ | 0.077 (0.054,0.095) | 0.111 (0.085,0.134) | 0.165 (0.134,0.209) | 0.328 (0.262,0.410) |
| $MSE_{ext}$ | 3.482 (3.266,3.794) | 3.557 (3.313,3.914) | 3.759 (3.436,4.214) | 4.799 (4.070,5.835) |
| $Optimism_{ext}$ | -0.719 (-1.300,-0.093) | -0.684 (-1.283,-0.050) | -0.683 (-1.281,0.017) | -0.860 (-1.706,-0.064) |
| $Optimism_{int}$ | -0.646 (-0.806,-0.521) | -0.578 (-0.731,-0.462) | -0.523 (-0.665,-0.420) | -0.486 (-0.620,-0.373) |
| $MSE_{corrected}$ | 3.262 (2.663,3.946) | 3.320 (2.716,4.034) | 3.474 (2.834,4.215) | 4.311 (3.534,5.291) |
| $\beta_{LP^*}$ | 1.025 (1.012,1.036) | 1.052 (1.035,1.069) | 1.092 (1.067,1.118) | 1.216 (1.167,1.273) |
| | **MCAR** | | | |
| $MSE_{apparent}$ | 3.013 (2.302,3.900) | 3.111 (2.347,4.052) | 3.331 (2.455,4.349) | 4.345 (3.109,5.696) |
| $\beta_{LP}$ | 1.030 (0.997,1.063) | 1.057 (1.012,1.096) | 1.102 (1.043,1.154) | 1.248 (1.155,1.349) |
| Tuning $\lambda$ | 0.046 (0.018,0.078) | 0.076 (0.028,0.120) | 0.131 (0.055,0.194) | 0.324 (0.196,0.447) |
| $MSE_{ext}$ | 3.956 (3.468,4.982) | 3.898 (3.473,4.668) | 3.994 (3.486,4.691) | 5.150 (3.885,6.541) |
| $Optimism_{ext}$ | -0.944 (-2.424,0.152) | -0.787 (-2.135,0.203) | -0.663 (-1.684,0.240) | -0.806 (-1.768,0.169) |
| $Optimism_{int}$ | -1.451 (-1.965,-1.047) | -1.286 (-1.734,-0.903) | -1.050 (-1.412,-0.744) | -0.678 (-0.917,-0.487) |
| $MSE_{corrected}$ | 4.464 (3.464,5.527) | 4.397 (3.566,5.254) | 4.381 (3.440,5.469) | 5.023 (3.571,6.411) |
| $\beta_{LP^*}$ | 0.940 (0.904,0.975) | 0.970 (0.930,1.003) | 1.012 (0.970,1.051) | 1.157 (1.088,1.219) |
| | **MAR** | | | |
| $MSE_{apparent}$ | 2.963 (2.156,3.693) | 3.078 (2.227,3.876) | 3.305 (2.387,4.181) | 4.303 (3.115,5.499) |
| $\beta_{LP}$ | 1.037 (1.012,1.065) | 1.065 (1.031,1.098) | 1.109 (1.063,1.151) | 1.251 (1.184,1.335) |
| Tuning $\lambda$ | 0.050 (0.023,0.079) | 0.082 (0.041,0.121) | 0.136 (0.079,0.192) | 0.325 (0.227,0.451) |
| $MSE_{ext}$ | 3.833 (3.418,4.588) | 3.811 (3.420,4.416) | 3.951 (3.495,4.584) | 5.128 (4.153,6.501) |
| $Optimism_{ext}$ | -0.870 (-2.083,-0.065) | -0.733 (-1.764,0.023) | -0.646 (-1.403,0.132) | -0.826 (-1.835,0.243) |
| $Optimism_{int}$ | -1.340 (-1.853,-1.012) | -1.170 (-1.595,-0.880) | -0.956 (-1.307,-0.696) | -0.645 (-0.883,-0.476) |
| $MSE_{corrected}$ | 4.304 (3.244,5.425) | 4.248 (3.198,5.372) | 4.260 (3.213,5.397) | 4.947 (3.696,6.182) |
| $\beta_{LP^*}$ | 0.955 (0.920,0.988) | 0.983 (0.948,1.020) | 1.027 (0.983,1.069) | 1.168 (1.100,1.235) |

Figure 2.2: **Optimism-corrected MSE** estimates from **MICE-Lasso** (ML) run on 300 simulated **20-covariate** datasets with 250 and 1000 observations (top and bottom rows respectively) comparing the scenarios with moderation assumption **S3** (without missing data), **S4** (with missing data, complete outcome), **S5** ( with missing data also in the outcome) and **S6** (missing data, complete outcome and interaction terms in the imputation model). ML estimated MSEs are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S3, the Lasso (L) corrected MSEs are shown.

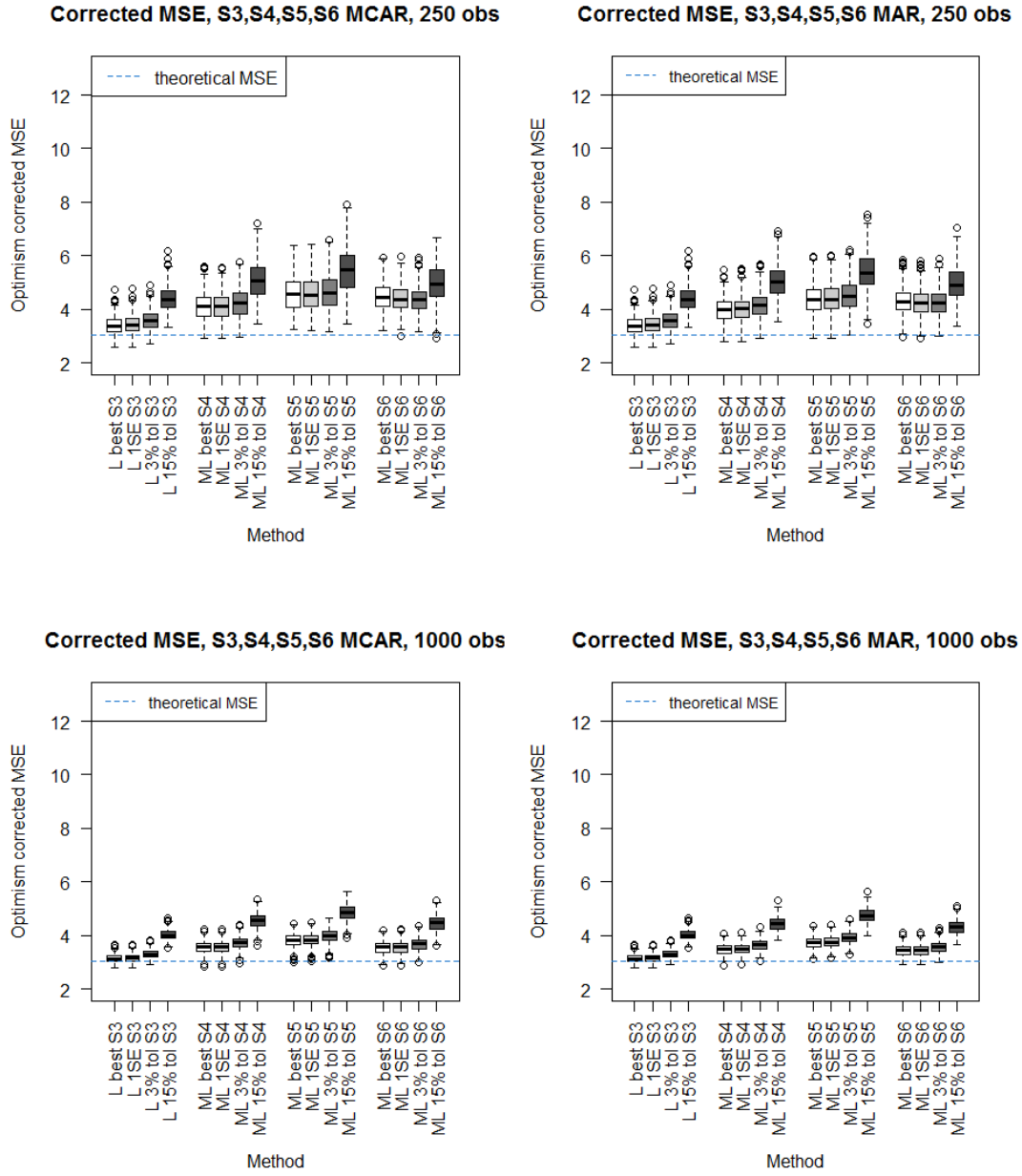


Figure 2.3: **Calibration slope** $\beta_{LP}$ estimates for **MICE-Lasso** (ML) run on 300 simulated **20-covariate** datasets with 250 and 1000 observations (top and bottom rows respectively) for the scenarios with moderation assumption **S3** (without missing data), **S4** (with missing data, complete outcome), **S5** ( with missing data also in the outcome) and **S6** (missing data, complete outcome and interaction terms in the imputation model). ML estimated calibration slopes are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S3, the Lasso (L) calibration slopes are shown.

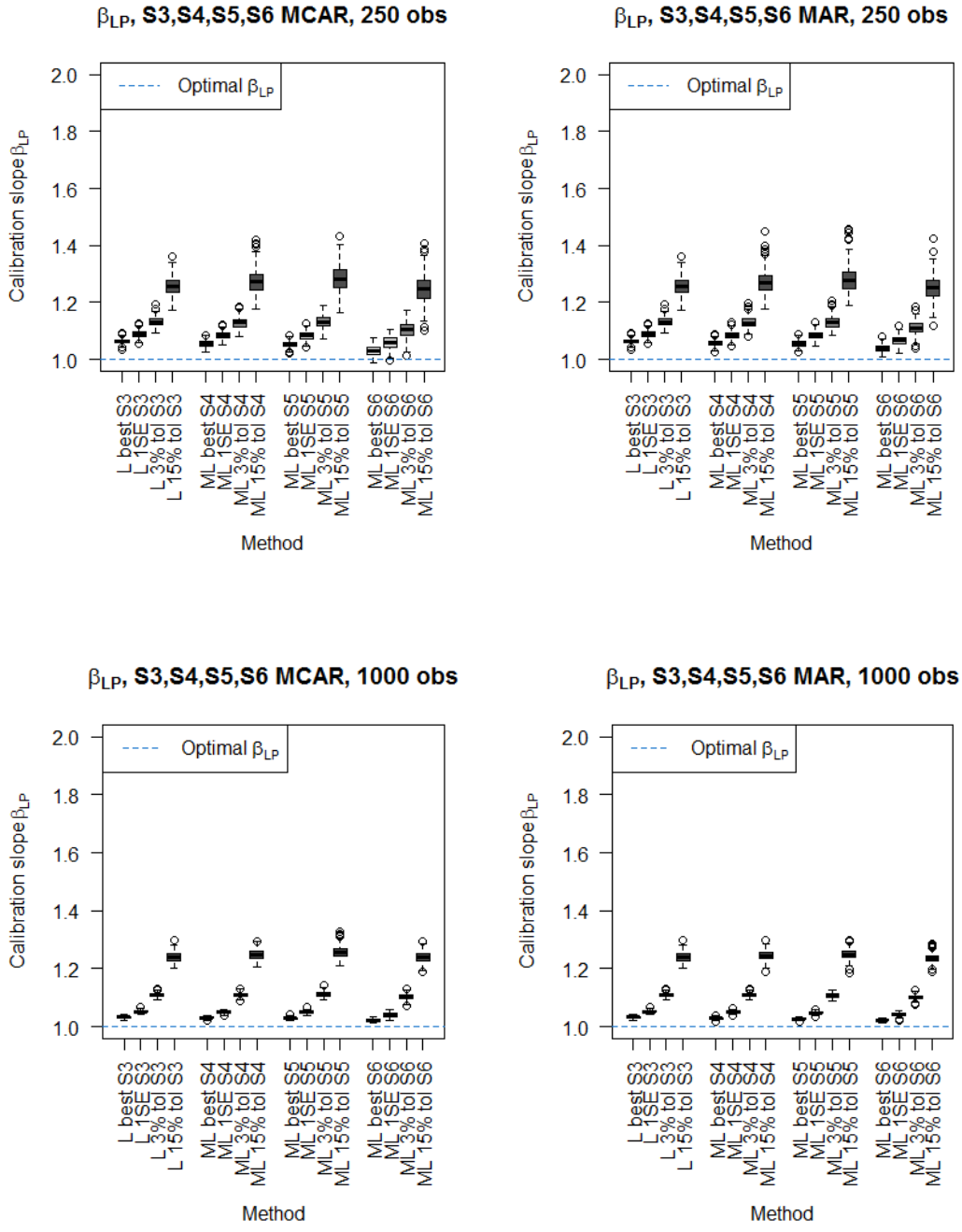


Figure 2.4: Average **internal and external MSE optimism** estimates with 2.5th and 97.5th percentiles for **MICE-Lasso** (ML) run on 300 simulated **20-covariate** datasets with 250 and 1000 observations for the scenarios with moderation assumption **S3** (without missing data), **S4** (with missing data, complete outcome), **S5** ( with missing data also in the outcome) and **S6** (missing data, complete outcome and interaction terms in the imputation model). ML estimated internal and external MSE optimism are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S3, the Lasso (L) optimism estimates are shown.

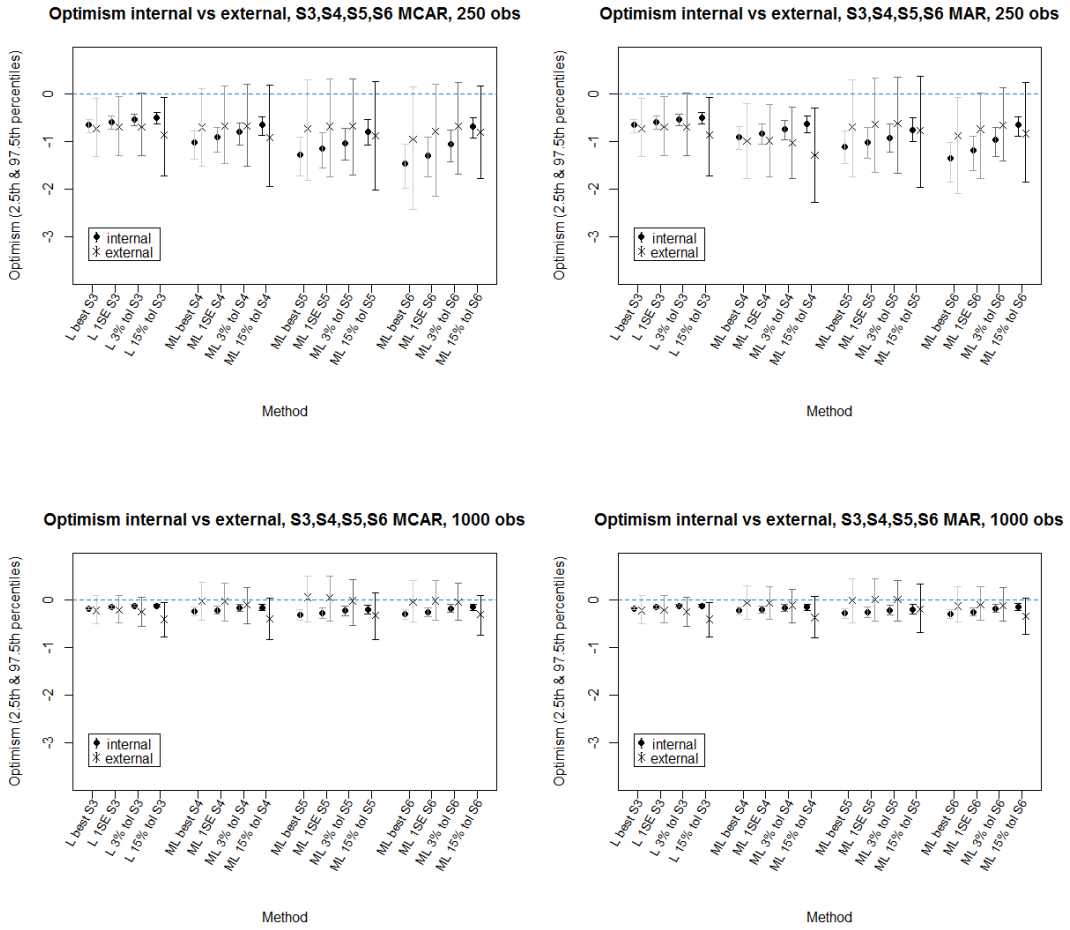


Figure 2.5: Average percentage of **true predictors (TP) selected among the actual TP** (SEN) estimates with 2.5th and 97.5th percentiles from **MICE-Lasso** (ML) run on 300 simulated **20-covariate** datasets with 250 and 1000 observations for the scenarios with moderation assumption **S3** (without missing data), **S4** (with missing data, complete outcome), **S5** ( with missing data also in the outcome) and **S6** (missing data, complete outcome and interaction terms in the imputation model). ML estimated percentages of TP selected among the actual TP variables are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S3, the Lasso (L) estimates are shown.

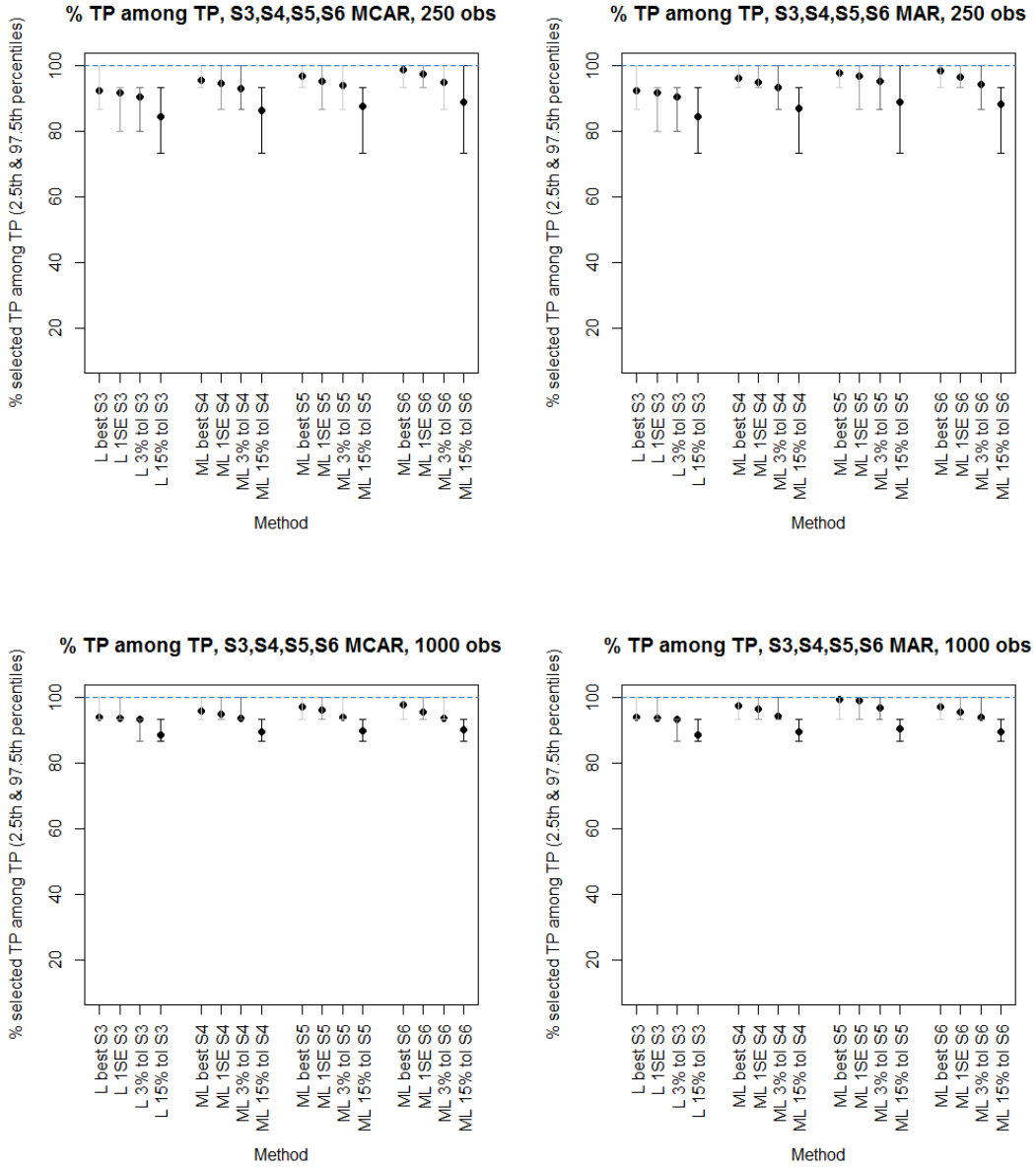


Figure 2.6: Average percentage of **true predictors (TP) among the selected variables** (PPV) estimates with 2.5th and 97.5th percentiles from **MICE-Lasso** (ML) run on 300 simulated **20-covariate** datasets with 250 and 1000 observations for the scenarios with moderation assumption **S3** (without missing data), **S4** (with missing data, complete outcome), **S5** ( with missing data also in the outcome) and **S6** (missing data, complete outcome and interaction terms in the imputation model). ML estimated percentages of TP among the selected variables are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S3, the Lasso (L) estimates are shown.

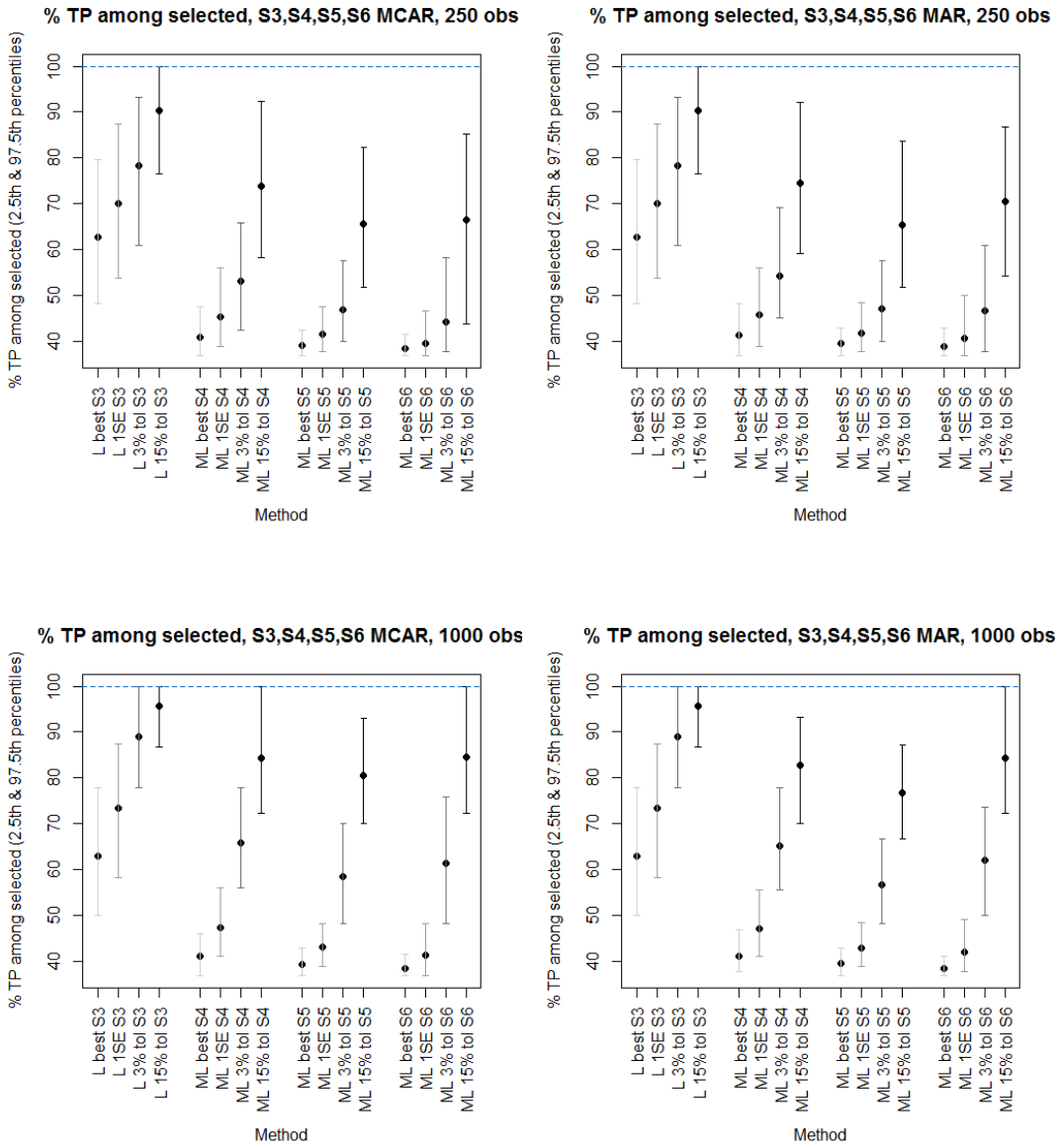
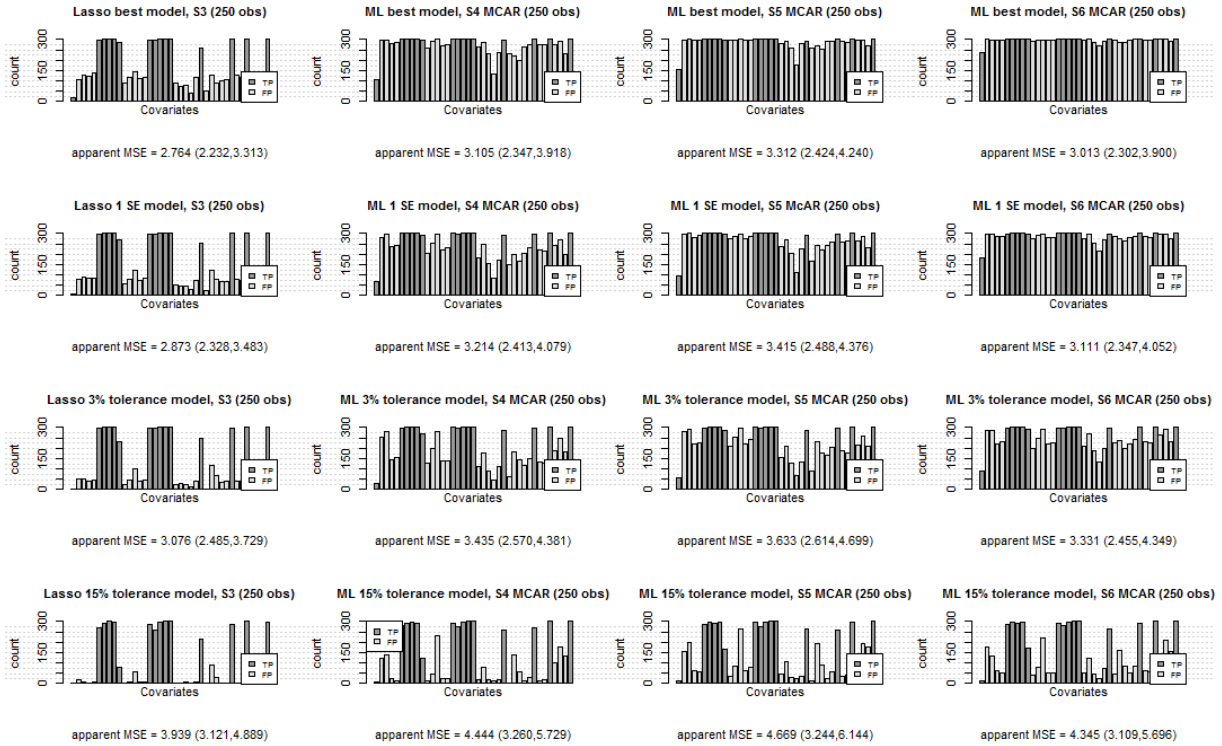


Figure 2.7: Comparison of **variable inclusion frequency** by **MICE-Lasso** (ML) run on 300 simulated **20-covariate** datasets with 250 observations for the scenarios with moderation assumption **S3** (without missing data), **S4** (with missing data, complete outcome), **S5** ( with missing data also in the outcome) and **S6** (missing data, complete outcome and interaction terms in the imputation model) with **MCAR** data. ML variable inclusion frequencies are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning.



**MissForest-Lasso: 20-covariate data results**

**MissForest-Lasso S2: Missing data, No assumption of moderation, complete outcome** When interactions were not included in the linear predictor, the best MissForest-Lasso model had an acceptable prediction accuracy performance similar to MICE-Lasso with a slightly smaller percentile interval ($MSE_{corrected}$=3.668 with 2.5th and 97.5th percentiles being 2.924 and 4.557 for MCAR, and $MSE_{corrected}$=3.610 with 2.5th and 97.5th percentiles being 2.886 and 4.346 for MAR 250 observation data, see Table 2.18). With increasing penalty tolerance, the $MSE_{corrected}$ was still acceptable apart from the 15% tolerance penalty. Internal and external MSE optimism estimates were similar in magnitude to Lasso's estimates in scenario S1 but the bias due to resampling was smaller for MissForest-Lasso than Lasso. The missing data imputation by MissForest seemed to adjust the optimism in the internal validation so that it was more similar to the optimism one would get by applying the model on completely new data. Also average calibration slope estimates were similar to Lasso scenario S1 for best and tolerance models, whilst MICE-Lasso (scenario S2) had slightly smaller estimates due to the poor variable selection leading to overfitting.

MissForest-Lasso had a variable inclusion frequency comparable to Lasso in scenario S1 (see Figures 2.29 and 2.28). The estimated PPV ranged from 64.9% (SD 8.4) for the best model to 96.8% (SD 5.9) for the 15% tolerance model for MCAR data and from 65.0% (SD 8.5) to 96.7% (SD 5.8) for MAR data; while the estimated SEN ranged from 99.4% (SD 2.4) to 75.2% (SD 15.7), with the tolerance models showing good variable selection, which was a result very similar to Lasso's results (see Table 2.17). MissForest-Lasso could also select up to 7% true models with 3% tolerance penalty for MCAR data and up to 8% for MAR data. The percentage of true models but one TP was estimated up to 24.7% for MCAR data and up to 21.7% for MAR data, much better performance compared to MICE-Lasso scenario S2 (see Table 2.3).

Results improved with larger sample size (see tables 2.17 and A.12 in the Appendix).

**Minor analysis 4 results** In scenario S2, MissForest-Lasso with 10 MissForest imputations outperformed MICE-Lasso in variable selection, suggesting the superiority of the imputation method MissForest compared to MICE (see Figure A.4 in the Appendix, comparing the single imputation MissForest-Lasso with the 10-imputation MissForest-Lasso).

Table 2.17: **Variable selection** simulation study results for scenarios **S1** (without missing data, no assumption of moderation) and **S2** (with missing data, complete outcome, no assumption of moderation) for **Lasso** and **MissForest-Lasso** best and tolerance models in the case of **20 covariates** and 300 samples of 250 and 1000 observations. The following results are shown: the estimated percentages of the times all the 10 true predictors (TP) are selected at the same time, the estimated percentages of the times the true model is selected apart from one variable, the percentages of the times the TP are the top ranked variables among the selected, the percentage of times the TP are the top ranked variables apart from one TP, the average percentages of selected TP among the TP (sensitivity, SEN), the average percentages of selected false positive predictors (FP) among FP (false positive rate, FPR), and the average percentages of selected TP among the selected variables (positive predictive value, PPV). The mean percentages are shown along with their standard deviation (SD)

| | **LASSO** | | | | **MissForest-LASSO** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Variable selection** | **Complete data** | | | | **MCAR** | | | | **MAR** | | | |
| | Best | 1 SE tolerance | 3% tolerance | 15% tolerance | Best | 1 SE tolerance | 3% tolerance | 15% tolerance | Best | 1 SE tolerance | 3% tolerance | 15% tolerance |
| | | | | | | **250 observations** | | | | | | |
| **% true models** | 0.7 | 3.7 | 16.0 | 13.0 | 0 | 1.0 | 7.0 | 5.0 | 0 | 2.3 | 9.0 | 2.0 |
| **% true models but 1** | 1.7 | 11.0 | 32.0 | 47.7 | 1.3 | 4.3 | 24.7 | 18.7 | 1.7 | 7.7 | 21.7 | 19.3 |
| **% TP in top 10** | 58.7 | 62.3 | 60.0 | 14.7 | 31.7 | 36.7 | 32.7 | 6.7 | 34.0 | 35.0 | 33.0 | 2.7 |
| **% TP in top 10 but 1** | 36.3 | 33.0 | 33.7 | 51.0 | 64.0 | 58.3 | 60.3 | 38.0 | 63.0 | 61.7 | 61.3 | 41.0 |
| **SEN (SD)** | 99.8 (1.4) | 99.5 (2.3) | 98.0 (4.0) | 87.6 (8.9) | 99.4 (2.4) | 98.4 (3.9) | 95.7 (6.3) | 75.2 (15.7) | 99.4 (2.3) | 98.8 (3.4) | 96.2 (5.9) | 76.2 (14.5) |
| **FPR (SD)** | 52.9 (18.7) | 33.4 (17.6) | 15.6 (12.5) | 1.8 (4.6) | 56.4 (19.9) | 37.6 (17.4) | 19.1 (13.4) | 2.8 (5.1) | 56.0 (19.1) | 36.7 (18.0) | 18.9 (13.4) | 3.1 (5.5) |
| **PPV (SD)** | 66.4 (8.7) | 76.2 (10.1) | 87.3 (9.0) | 98.2 (4.4) | 64.9 (8.4) | 73.5 (9.2) | 84.4 (9.3) | 96.8 (5.9) | 65.0 (8.5) | 74.2 (9.9) | 84.6 (9.4) | 96.7 (5.8) |
| | | | | | | **1000 observations** | | | | | | |
| **% true models** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **% true models but 1** | 0 | 0.3 | 13.7 | 15.7 | 0 | 0 | 0 | 1.0 | 0 | 0 | 0 | 1.0 |
| **% TP in top 10** | 1.3 | 1.7 | 13.7 | 15.7 | 0.7 | 0.7 | 0 | 1.0 | 1.3 | 0.7 | 0 | 1.0 |
| **% TP in top 10 but 1** | 89.3 | 90.3 | 76.3 | 46.7 | 34.7 | 51.7 | 65.0 | 25.0 | 48.3 | 62.7 | 67.0 | 24.3 |
| **SEN (SD)** | 93.7 (1.8) | 93.5 (1.3) | 93.2 (1.3) | 88.6 (3.2) | 97.5 (3.2) | 95.5 (3.1) | 93.7 (1.6) | 89.9 (3.4) | 97.0 (3.3) | 95.5 (3.1) | 93.7 (1.6) | 89.5 (3.4) |
| **FPR (SD)** | 35.7 (11.1) | 22.1 (9.1) | 7.5 (5.0) | 2.6 (3.2) | 97.4 (4.1) | 85.4 (10.3) | 38.2 (11.5) | 10.7 (5.6) | 96.7 (4.0) | 83.3 (9.6) | 36.9 (10.5) | 10.9 (5.5) |
| **PPV (SD)** | 63.0 (7.5) | 73.5 (8.2) | 89.1 (6.6) | 95.7 (5.0) | 38.5 (1.3) | 41.3 (3.1) | 61.4 (7.1) | 84.6 (7.1) | 38.5 (1.2) | 41.9 (2.9) | 62.1 (6.6) | 84.3 (6.9) |

Table 2.18: **Accuracy** simulation study results for **MissForest-Lasso** analysis with Harrell bootstrap validation: scenarios **S1** (without missing data, no assumption of moderation) and **S2** (with missing data, complete outcome, no assumption of moderation) based on 300 data sets of **20 variables** each (**n=250**). Means of all estimates along with their corresponding 2.5th and 97.5th percentile values within parenthesis. The theoretical MSE is $1.74^2 = 3.028$

| **Estimates** | **Complete cases** | | | |
|---|---|---|---|---|
| | Best model | 1 SE tolerance | 3% tolerance | 15% tolerance |
| $MSE_{apparent}$ | 2.840 (2.315,3.406) | 2.937 (2.398,3.544) | 3.127 (2.553,3.772) | 4.002 (3.271,4.905) |
| $\beta_{LP}$ | 1.062 (1.041,1.084) | 1.102 (1.075,1.131) | 1.161 (1.124,1.200) | 1.334 (1.261,1.429) |
| Tuning $\lambda$ | 0.065 (0.039,0.085) | 0.106 (0.076,0.134) | 0.168 (0.134,0.209) | 0.360 (0.293,0.410) |
| $MSE_{ext}$ | 3.509 (3.289,3.791) | 3.590 (3.347,3.890) | 3.781 (3.486,4.172) | 4.718 (4.113,5.420) |
| $Optimism_{ext}$ | -0.669 (-1.212,-0.010) | -0.653 (-1.198,0.014) | -0.654 (-1.216,0.055) | -0.716 (-1.460,0.176) |
| $Optimism_{int}$ | -0.422 (-0.534,-0.333) | -0.383 (-0.482,-0.295) | -0.346 (-0.442,-0.267) | -0.309 (-0.404,-0.216) |
| $MSE_{corrected}$ | 3.262 (2.663,3.946) | 3.320 (2.716,4.034) | 3.474 (2.834,4.215) | 4.311 (3.534,5.291) |
| $\beta_{LP^*}$ | 1.028 (1.015,1.042) | 1.069 (1.051,1.088) | 1.126 (1.100,1.157) | 1.296 (1.235,1.379) |
| | **MCAR** | | | |
| $MSE_{apparent}$ | 3.156 (2.583,3.900) | 3.271 (2.672,4.037) | 3.503 (2.843,4.279) | 4.574 (3.594,5.664) |
| $\beta_{LP}$ | 1.065 (1.041,1.089) | 1.107 (1.079,1.140) | 1.171 (1.130,1.214) | 1.368 (1.278,1.473) |
| Tuning $\lambda$ | 0.069 (0.041,0.095) | 0.115 (0.081,0.149) | 0.187 (0.149,0.234) | 0.429 (0.328,0.574) |
| $MSE_{ext}$ | 3.764 (3.417,4.289) | 3.855 (3.479,4.381) | 4.098 (3.597,4.739) | 5.276 (4.416,6.221) |
| $Optimism_{ext}$ | -0.608 (-1.287,0.121) | -0.584 (-1.262,0.136) | -0.595 (-1.323,0.162) | -0.702 (-1.601,0.259) |
| $Optimism_{int}$ | -0.512 (-0.717,-0.329) | -0.459 (-0.649,-0.284) | -0.411 (-0.584,-0.247) | -0.341 (-0.481,-0.200) |
| $MSE_{corrected}$ | 3.668 (2.924,4.557) | 3.730 (2.971,4.629) | 3.914 (3.109,4.818) | 4.915 (3.910,6.027) |
| $\beta_{LP^*}$ | 1.023 (1.002,1.044) | 1.068 (1.043,1.093) | 1.133 (1.099,1.169) | 1.335 (1.264,1.431) |
| | **MAR** | | | |
| $MSE_{apparent}$ | 3.109 (2.465,3.772) | 3.221 (2.550,3.926) | 3.446 (2.732,4.187) | 4.500 (3.600,5.612) |
| $\beta_{LP}$ | 1.064 (1.039,1.089) | 1.106 (1.075,1.139) | 1.169 (1.129,1.217) | 1.368 (1.280,1.494) |
| Tuning $\lambda$ | 0.069 (0.035,0.095) | 0.115 (0.076,0.149) | 0.186 (0.134,0.234) | 0.428 (0.328,0.574) |
| $MSE_{ext}$ | 3.788 (3.396,4.420) | 3.879 (3.460,4.523) | 4.112 (3.600,4.756) | 5.262 (4.429,6.206) |
| $Optimism_{ext}$ | -0.678 (-1.541,0.107) | -0.657 (-1.487,0.124) | -0.667 (-1.507,0.115) | -0.762 (-1.673,0.169) |
| $Optimism_{int}$ | -0.501 (-0.714,-0.329) | -0.449 (-0.640,-0.287) | -0.401 (-0.573,-0.250) | -0.337 (-0.479,-0.200) |
| $MSE_{corrected}$ | 3.610 (2.886,4.346) | 3.670 (2.929,4.429) | 3.847 (3.088,4.641) | 4.837 (3.938,5.964) |
| $\beta_{LP^*}$ | 1.025 (0.999,1.049) | 1.070 (1.045,1.099) | 1.133 (1.104,1.174) | 1.335 (1.266,1.448) |

**MissForest-Lasso S4: Missing data, Assumption of moderation, complete outcome** Discrimination performance for MissForest-Lasso in presence of missing data and moderators among the predictors was comparable to MICE-Lasso in the same scenario. The optimism-corrected MSE for the best model was poor compared to the theoretical MSE: 4.174 (2.5th and 97.5th percentiles: 3.360, 5.101) for MCAR data and slightly less 4.087 (2.5th and 97.5th percentiles: 3.292, 5.094) for MAR data with 250 observations (see Table 2.20). The estimates of internal and external MSE optimism were again the smallest in absolute value after Lasso and Elasticnet. However, the difference between internal and external MSE optimism were not improved compared to MICE-Lasso. MissForest-Lasso behaved similarly with MAR and MCAR data, denoting that the RF imputation algorithm was less sensitive to correlation between variables. Calibration slopes estimates were good, also close to MICE-Lasso estimates.

Variable inclusion frequency for MissForest-Lasso is again similar to Lasso in scenario S3 with complete data (see Figures 2.30, 2.33 and 2.31). The TPs were almost always selected: average percentage of TP among TPs was 90.2 (SD 4.4) for the best model for MCAR data to a minimum of 71.8 (SD 10.8) for the 15% tolerance model (slightly better for MAR data, see 2.19). Among the selected variables on average 62.8% (SD 7.3) were TPs for the best model and 89.9% (SD 7.2) for the 15% tolerance model. Therefore, only the 3% and 15% tolerance models showed good variable selection.

The 1000 observation dataset MissForest-Lasso results were even closer to Lasso results (see tables 2.19, A.13 in the Appendix).

Table 2.19: **Variable selection** simulation study results for scenarios **S3** (assumption of moderation, without missing data) and **S4** (assumption of moderation, with missing data) for **MissForest-Lasso** best and tolerance models in the case of **20 covariates** and 300 samples of 250 and 1000 observations. The following results are shown: the estimated percentages of the times all the 10 true predictors (TP) are selected at the same time, the estimated percentages of the times the true model is selected apart from one variable, the percentages of the times the TP are the top ranked variables among the selected, the percentage of times the TP are the top ranked variables apart from one TP, the average percentages of selected TP among the TP (sensitivity, SEN), the average percentages of selected false positive predictors (FP) among FP (false positive rate, FPR), and the average percentages of selected TP among the selected variables (positive predictive value, PPV). The mean percentages are shown along with their standard deviation (SD)

| | **LASSO** | | | | **MissForest-LASSO** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Variable selection** | **Complete data** | | | | **MCAR** | | | | **MAR** | | | |
| | Best | 1 SE tolerance | 3% tolerance | 15% tolerance | Best | 1 SE tolerance | 3% tolerance | 15% tolerance | Best | 1 SE tolerance | 3% tolerance | 15% tolerance |
| | | | | | | **250 observations** | | | | | | |
| **% true models** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **% true models but 1** | 0 | 0 | 1.3 | 2.7 | 0 | 0 | 0.7 | 1.0 | 0 | 0 | 0.7 | 0 |
| **% TP in top 10** | 0 | 0 | 1.0 | 2.7 | 0 | 0 | 0.7 | 1.0 | 0 | 0 | 0.7 | 0 |
| **% TP in top 10 but 1** | 17.3 | 20.3 | 20.7 | 13.7 | 5.3 | 8.0 | 8.0 | 3.0 | 5.0 | 5.0 | 5.3 | 2.7 |
| **SEN (SD)** | 92.3 (3.3) | 91.7 (3.5) | 90.5 (4.2) | 84.2 (6.0) | 90.2 (4.4) | 89.1 (4.8) | 86.3 (6.2) | 71.8 (10.8) | 91.0 (4.5) | 90.2 (4.6) | 87.5 (5.7) | 73.5 (9.9) |
| **FPR (SD)** | 35.9 (11.9) | 25.6 (10.2) | 16.6 (7.9) | 5.9 (4.6) | 36.7 (10.8) | 26.3 (9.1) | 16.8 (8.0) | 5.4 (4.3) | 37.6 (12.4) | 27.6 (10.9) | 17.4 (8.3) | 6.0 (4.6) |
| **PPV (SD)** | 62.6 (7.9) | 70.2 (8.6) | 78.2 (8.3) | 90.4 (6.7) | 62.8 (7.3) | 68.8 (7.8) | 77.2 (8.6) | 89.9 (7.2) | 62.4 (7.8) | 68.3 (8.7) | 76.9 (8.6) | 89.1 (7.6) |
| | | | | | | **1000 observations** | | | | | | |
| **% true models** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **% true models but 1** | 0 | 0.3 | 13.7 | 15.7 | 0 | 0 | 4.7 | 4.7 | 0 | 0.3 | 3.7 | 3.0 |
| **% TP in top 10** | 1.3 | 1.7 | 13.7 | 15.7 | 2.3 | 1.0 | 4.7 | 4.7 | 1.7 | 1.3 | 3.0 | 3.0 |
| **% TP in top 10 but 1** | 89.3 | 90.3 | 76.3 | 46.7 | 63.3 | 71.3 | 65.0 | 40.7 | 63.0 | 68.3 | 64.3 | 37.3 |
| **SEN (SD)** | 93.7 (1.8) | 93.5 (1.3) | 93.2 (1.3) | 88.6 (3.2) | 93.8 (2.3) | 93.5 (1.8) | 92.8 (1.8) | 86.1 (3.6) | 93.9 (2.4) | 93.6 (2.1) | 92.9 (2.0) | 86.4 (3.1) |
| **FPR (SD)** | 35.7 (11.1) | 22.1 (9.1) | 7.5 (5.0) | 2.6 (3.2) | 38.3 (10.9) | 24.2 (9.4) | 9.2 (5.3) | 2.7 (3.0) | 40.9 (11.7) | 27.2 (9.6) | 11.0 (5.9) | 3.5 (3.7) |
| **PPV (SD)** | 63.0 (7.5) | 73.5 (8.2) | 89.1 (6.6) | 95.7 (5.0) | 62.7 (6.7) | 71.6 (8.1) | 86.8 (6.7) | 95.5 (4.9) | 61.4 (6.9) | 69.1 (7.7) | 84.7 (7.0) | 94.3 (5.7) |

Table 2.20: **Accuracy** simulation study results for **MissForest-Lasso** analysis with Harrell bootstrap validation: scenarios **S3** (assumption of moderation, without missing data) and **S4** (assumption of moderation, with missing data) based on 300 data sets of **20 variables** each (**n=250**). Means of all estimates along with their corresponding 2.5th and 97.5th percentile values within parenthesis. The theoretical MSE is $1.74^2 = 3.028$

| **Estimates** | **Complete cases** | | | |
|---|---|---|---|---|
| | Best model | 1 SE tolerance | 3% tolerance | 15% tolerance |
| $\text{MSE}_{apparent}$ | 2.764 (2.232,3.313) | 2.873 (2.328,3.483) | 3.076 (2.485,3.729) | 3.939 (3.121,4.889) |
| $\beta_{LP}$ | 1.062 (1.044,1.079) | 1.089 (1.069,1.112) | 1.132 (1.102,1.165) | 1.259 (1.204,1.332) |
| Tuning $\lambda$ | 0.077 (0.054,0.095) | 0.111 (0.085,0.134) | 0.165 (0.134,0.209) | 0.328 (0.262,0.410) |
| $\text{MSE}_{ext}$ | 3.482 (3.266,3.794) | 3.557 (3.313,3.914) | 3.759 (3.436,4.214) | 4.799 (4.070,5.835) |
| $\text{Optimism}_{ext}$ | -0.719 (-1.300,-0.093) | -0.684 (-1.283,-0.050) | -0.683 (-1.281,0.017) | -0.860 (-1.706,-0.064) |
| $\text{Optimism}_{int}$ | -0.646 (-0.806,-0.521) | -0.578 (-0.731,-0.462) | -0.523 (-0.665,-0.420) | -0.486 (-0.620,-0.373) |
| $\text{MSE}_{corrected}$ | 3.410 (2.788,4.131) | 3.451 (2.826,4.194) | 3.598 (2.937,4.384) | 4.425 (3.541,5.495) |
| $\beta_{LP*}$ | 1.025 (1.012,1.036) | 1.052 (1.035,1.069) | 1.092 (1.067,1.118) | 1.216 (1.167,1.273) |
| | **MCAR** | | | |
| $\text{MSE}_{apparent}$ | 3.364 (2.697,4.191) | 3.508 (2.796,4.397) | 3.785 (3.001,4.746) | 4.991 (3.912,6.283) |
| $\beta_{LP}$ | 1.069 (1.049,1.091) | 1.099 (1.074,1.129) | 1.148 (1.110,1.190) | 1.310 (1.233,1.404) |
| Tuning $\lambda$ | 0.088 (0.068,0.513) | 0.129 (0.068,0.513) | 0.197 (0.068,0.513) | 0.426 (0.068,0.513) |
| $\text{MSE}_{ext}$ | 3.822 (3.427,4.364) | 3.958 (3.505,4.631) | 4.316 (3.692,5.154) | 5.959 (4.775,7.447) |
| $\text{Optimism}_{ext}$ | -0.458 (-1.177,0.356) | -0.450 (-1.212,0.446) | -0.531 (-1.390,0.422) | -0.968 (-2.099,0.219) |
| $\text{Optimism}_{int}$ | -0.825 (-1.073,-0.596) | -0.725 (-0.959,-0.519) | -0.643 (-0.840,-0.447) | -0.555 (-0.736,-0.382) |
| $\text{MSE}_{corrected}$ | 4.174 (3.360,5.101) | 4.218 (3.377,5.188) | 4.407 (3.527,5.440) | 5.519 (4.345,6.897) |
| $\beta_{LP*}$ | 1.021 (1.001,1.043) | 1.055 (1.033,1.082) | 1.103 (1.074,1.140) | 1.261 (1.207,1.335) |
| | **MAR** | | | |
| $\text{MSE}_{apparent}$ | 3.269 (2.633,4.077) | 3.409 (2.802,4.230) | 3.671 (2.977,4.573) | 4.838 (3.901,6.094) |
| $\beta_{LP}$ | 1.066 (1.043,1.087) | 1.096 (1.070,1.126) | 1.144 (1.114,1.181) | 1.303 (1.233,1.395) |
| Tuning $\lambda$ | 0.084 (0.061,0.459) | 0.125 (0.061,0.459) | 0.191 (0.061,0.459) | 0.414 (0.061,0.459) |
| $\text{MSE}_{ext}$ | 3.792 (3.418,4.341) | 3.919 (3.464,4.562) | 4.248 (3.611,5.112) | 5.843 (4.725,7.240) |
| $\text{Optimism}_{ext}$ | -0.524 (-1.353,0.239) | -0.510 (-1.308,0.299) | -0.577 (-1.441,0.291) | -1.005 (-2.240,0.109) |
| $\text{Optimism}_{int}$ | -0.815 (-1.085,-0.605) | -0.721 (-0.983,-0.520) | -0.644 (-0.890,-0.450) | -0.563 (-0.766,-0.393) |
| $\text{MSE}_{corrected}$ | 4.087 (3.292,5.094) | 4.132 (3.344,5.124) | 4.315 (3.466,5.338) | 5.395 (4.309,6.784) |
| $\beta_{LP^*}$ | 1.023 (1.001,1.044) | 1.056 (1.030,1.080) | 1.103 (1.072,1.138) | 1.258 (1.204,1.332) |

**MissForest-Lasso S5: Assumption of moderation, Missing data also in outcome (20% missingness MAR and MCAR)** When there were missing data also in the outcome variable, MissForest-Lasso outperformed MICE-Lasso in the same scenario with an acceptable optimism-corrected MSE of 3.704 (2.5th and 97.5th percentiles: 2.926 and 4.588) for the best model when the sample size was 250 and the data were MAR (see Table 2.22 for the worse but still acceptable MCAR data result). With increasing penalty tolerance, only the 1 SE and 3% corrected MSE of the MAR case were still acceptable, while the MCAR estimates were all very poor. MissForest-Lasso had better discrimination performance when the outcome had missing data compared to when it was complete (scenario S4) for these simulated data. However, there is a problem underlining this apparent improvement: while the internal MSE optimism estimates were smaller in absolute values than estimates in scenario S3, there was the largest difference between internal and external MSE optimism as the latter was large in absolute value (see Figure 2.16). As a consequence, this good model performance would not be likely to replicate on new observations. Calibration was similar to scenario S4 when the outcome was complete.

Variable selection for this scenario was worse than S4 (see Table 2.21 and figures 2.34 and 2.32). This time only the 3% tolerance model variable selection was good, the other models' performances were acceptable apart from the 15% tolerance model performance which was poor.

The results slightly improved when the sample size was 1000, but the performance was still inferior compared to other scenarios (see tables A.14 and 2.21).

Table 2.21: **Variable selection** simulation study results for scenarios **S3** (assumption of moderation, without missing data) and **S5** (assumption of moderation, with missing data also in the outcome) for **Lasso** and **MissForest-Lasso** best and tolerance models in the case of **20 covariates** and 300 samples of 250 and 1000 observations. The following results are shown: the estimated percentages of the times all the 10 true predictors (TP) are selected at the same time, the estimated percentages of the times the true model is selected apart from one variable, the percentages of the times the TP are the top ranked variables among the selected, the percentage of times the TP are the top ranked variables apart from one TP, the average percentages of selected TP among the TP (sensitivity, SEN), the average percentages of selected false positive predictors (FP) among FP (false positive rate, FPR), and the average percentages of selected TP among the selected variables (positive predictive value, PPV). The mean percentages are shown along with their standard deviation (SD)

| | **LASSO** | | | | **MissForest-LASSO** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Variable selection** | **Complete data** | | | | **MCAR** | | | | **MAR** | | | |
| | Best | 1 SE tolerance | 3% tolerance | 15% tolerance | Best | 1 SE tolerance | 3% tolerance | 15% tolerance | Best | 1 SE tolerance | 3% tolerance | 15% tolerance |
| | | | | | | **250 observations** | | | | | | |
| **% true models** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **% true models but 1** | 0 | 0 | 1.3 | 2.7 | 0 | 0 | 0 | 0.3 | 0 | 0 | 0 | 0 |
| **% TP in top 10** | 0 | 0 | 1.0 | 2.7 | 0 | 0 | 0 | 0.3 | 0 | 0 | 0 | 0 |
| **% TP in top 10 but 1** | 17.3 | 20.3 | 20.7 | 13.7 | 0 | 0.3 | 0.7 | 0.3 | 2.7 | 1.7 | 2.3 | 0.7 |
| **SEN (SD)** | 92.3 (3.3) | 91.7 (3.5) | 90.5 (4.2) | 84.2 (6.0) | 88.6 (5.3) | 86.0 (6.3) | 82.1 (7.6) | 62.0 (12.3) | 90.2 (5.0) | 88.4 (5.9) | 85.4 (6.4) | 67.9 (11.8) |
| **FPR (SD)** | 35.9 (11.9) | 25.6 (10.2) | 16.6 (7.9) | 5.9 (4.6) | 36.0 (11.1) | 25.8 (10.5) | 16.4 (7.8) | 4.8 (4.4) | 40.1 (12.1) | 29.9 (10.6) | 20.5 (8.8) | 6.7 (5.0) |
| **PPV (SD)** | 62.6 (7.9) | 70.2 (8.6) | 78.2 (8.3) | 90.4 (6.7) | 62.9 (7.2) | 68.8 (8.7) | 76.8 (8.8) | 89.9 (8.4) | 61.7 (7.8) | 65.8 (8.2) | 73.3 (8.6) | 87.2 (8.4) |
| | | | | | | **1000 observations** | | | | | | |
| **% true models** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **% true models but 1** | 0 | 0.3 | 13.7 | 15.7 | 0 | 0 | 4.3 | 0.7 | 0 | 0 | 0.7 | 0 |
| **% TP in top 10** | 1.3 | 1.7 | 13.7 | 15.7 | 1.3 | 1.0 | 4.0 | 0.7 | 1.3 | 1.3 | 0.7 | 0 |
| **% TP in top 10 but 1** | 89.3 | 90.3 | 76.3 | 46.7 | 51.7 | 55.0 | 42.7 | 24.3 | 47.7 | 47.0 | 28.0 | 19.0 |
| **SEN (SD)** | 93.7 (1.8) | 93.5 (1.3) | 93.2 (1.3) | 88.6 (3.2) | 93.8 (2.6) | 93.4 (2.1) | 92.2 (2.8) | 83.8 (4.6) | 94.2 (2.5) | 93.8 (2.3) | 91.8 (3.0) | 85.6 (3.1) |
| **FPR (SD)** | 35.7 (11.1) | 22.1 (9.1) | 7.5 (5.0) | 2.6 (3.2) | 40.2 (11.2) | 26.1 (9.5) | 10.8 (5.4) | 3.4 (3.3) | 46.7 (11.9) | 33.3 (10.2) | 15.2 (7.0) | 5.3 (4.0) |
| **PPV (SD)** | 63.0 (7.5) | 73.5 (8.2) | 89.1 (6.6) | 95.7 (5.0) | 61.8 (7.0) | 70.0 (8.1) | 84.8 (6.8) | 94.3 (5.4) | 59.3 (6.7) | 64.6 (7.4) | 79.8 (7.4) | 91.3 (6.1) |

Table 2.22: **Accuracy** simulation study results for **MissForest-Lasso** analysis with Harrell bootstrap validation: scenarios **S3** (assumption of moderation, complete data) and **S5** (assumption of moderation, missing data also in the outcome), based on 300 data sets of **20 variables** each (**n=250**). Means of all estimates along with their corresponding 2.5th and 97.5th percentile values within parenthesis. The theoretical MSE is $1.74^2 = 3.028$

| **Estimates** | **Complete cases** | | | |
|---|---|---|---|---|
| | Best model | 1 SE tolerance | 3% tolerance | 15% tolerance |
| $\text{MSE}_{apparent}$ | 2.764 (2.232,3.313) | 2.873 (2.328,3.483) | 3.076 (2.485,3.729) | 3.939 (3.121,4.889) |
| $\beta_{LP}$ | 1.062 (1.044,1.079) | 1.089 (1.069,1.112) | 1.132 (1.102,1.165) | 1.259 (1.204,1.332) |
| Tuning $\lambda$ | 0.077 (0.054,0.095) | 0.111 (0.085,0.134) | 0.165 (0.134,0.209) | 0.328 (0.262,0.410) |
| $\text{MSE}_{ext}$ | 3.482 (3.266,3.794) | 3.557 (3.313,3.914) | 3.759 (3.436,4.214) | 4.799 (4.070,5.835) |
| $\text{Optimism}_{ext}$ | -0.719 (-1.300,-0.093) | -0.684 (-1.283,-0.050) | -0.683 (-1.281,0.017) | -0.860 (-1.706,-0.064) |
| $\text{Optimism}_{int}$ | -0.646 (-0.806,-0.521) | -0.578 (-0.731,-0.462) | -0.523 (-0.665,-0.420) | -0.486 (-0.620,-0.373) |
| $\text{MSE}_{corrected}$ | 3.410 (2.788,4.131) | 3.451 (2.826,4.194) | 3.598 (2.937,4.384) | 4.425 (3.541,5.495) |
| $\beta_{LP*}$ | 1.025 (1.012,1.036) | 1.052 (1.035,1.069) | 1.092 (1.067,1.118) | 1.216 (1.167,1.273) |
| | **MCAR** | | | |
| $\text{MSE}_{apparent}$ | 3.132 (2.448,3.872) | 3.283 (2.553,4.032) | 3.551 (2.764,4.388) | 4.761 (3.686,5.978) |
| $\beta_{LP}$ | 1.075 (1.050,1.102) | 1.108 (1.077,1.146) | 1.161 (1.118,1.216) | 1.343 (1.255,1.470) |
| Tuning $\lambda$ | 0.090 (0.068,0.513) | 0.134 (0.068,0.513) | 0.204 (0.068,0.513) | 0.455 (0.068,0.513) |
| $\text{MSE}_{ext}$ | 4.365 (3.743,5.132) | 4.619 (3.895,5.522) | 5.106 (4.163,6.226) | 7.083 (5.582,8.833) |
| $\text{Optimism}_{ext}$ | -1.234 (-2.021,-0.334) | -1.336 (-2.189,-0.380) | -1.555 (-2.575,-0.506) | -2.323 (-3.692,-0.931) |
| $\text{Optimism}_{int}$ | -0.831 (-1.122,-0.582) | -0.737 (-0.979,-0.522) | -0.659 (-0.882,-0.461) | -0.547 (-0.760,-0.372) |
| $\text{MSE}_{corrected}$ | 3.963 (3.094,4.937) | 4.019 (3.108,5.013) | 4.210 (3.265,5.200) | 5.308 (4.175,6.583) |
| $\beta_{LP*}$ | 1.028 (1.004,1.050) | 1.066 (1.039,1.097) | 1.118 (1.084,1.160) | 1.297 (1.220,1.399) |
| | **MAR** | | | |
| $\text{MSE}_{apparent}$ | 2.910 (2.233,3.653) | 3.046 (2.351,3.860) | 3.288 (2.519,4.152) | 4.402 (3.417,5.579) |
| beta LP | 1.068 (1.046,1.094) | 1.100 (1.074,1.135) | 1.149 (1.113,1.194) | 1.321 (1.241,1.450) |
| $\beta_{LP}$ | 1.068 (1.046,1.094) | 1.100 (1.074,1.135) | 1.149 (1.113,1.194) | 1.321 (1.241,1.450) |
| Tuning $\lambda$ | 0.081 (0.061,0.513) | 0.122 (0.061,0.513) | 0.185 (0.061,0.513) | 0.415 (0.061,0.513) |
| $\text{MSE}_{ext}$ | 4.260 (3.637,5.099) | 4.461 (3.738,5.452) | 4.872 (3.961,6.137) | 6.680 (5.168,8.522) |
| $\text{Optimism}_{ext}$ | -1.351 (-2.478,-0.554) | -1.415 (-2.543,-0.582) | -1.584 (-2.816,-0.624) | -2.278 (-3.766,-0.978) |
| $\text{Optimism}_{int}$ | -0.795 (-1.073,-0.570) | -0.711 (-0.960,-0.507) | -0.645 (-0.860,-0.447) | -0.559 (-0.744,-0.374) |
| $\text{MSE}_{corrected}$ | 3.704 (2.926,4.588) | 3.757 (2.959,4.669) | 3.933 (3.087,4.874) | 4.961 (3.935,6.081) |
| $\beta_{LP^*}$ | 1.027 (1.002,1.053) | 1.063 (1.034,1.095) | 1.112 (1.075,1.155) | 1.281 (1.212,1.379) |

**MissForest-Random Forests: 20-covariate data results**

**Random Forests S1: No missing data, No assumption of moderation** RF discrimination performance was the worst among the methods in scenario S1 without missing data: average $MSE_{corrected}$=5.014 (2.5th and 97.5th percentiles 4.194 and 5.915, see Table 2.24) for a corrected pseudo-$R^2$ of 0.523 (2.5th and 97.5th percentiles 0.442 and 0.599) in the 250 observation analysis. However, there was no bias due to internal validation resampling, meaning that this performance would reflect the performance the model will have on new data.

RF tended to give more importance to continuous variables than binary ones even when they were FP (see Figure 2.27). As a results, TPs had variable an importance rank frequency similar to the inclusion frequency of the 15% tolerance model of Elasticnet, but continuous FP were selected more times that binary FP (see Figure 2.27: the first 10 variables are binary and the last 10 are continuous). Only 2.3% of the times RF had the TPs as the 10 most important variables (see Table 2.23), whilst Lasso had them in the top ten 58.7% (see Table 2.3) and Elasticnet 56.7% (see Table 2.5) for the best models.

When the sample size was 1000, all results improved a great deal as expected even though the model was still underfitting the data (see tables 2.23 and 2.24).

Table 2.23: **Variable selection** simulation study results for scenarios **S1** (without missing data, no assumption of moderation) and **S2** (with missing data, complete outcome, no assumption of moderation) for **Random Forest** and **MissForest-Random Forests** (MR) best models in the case of **20 covariates** and 300 samples of 250 and 1000 observations. The following results are shown: the percentages of the times the TP are the top ranked variables among the selected, the percentage of times the TP are the top ranked variables apart from one TP.

| | MissForest-RANDOM FOREST | | | | | |
|---|---|---|---|---|---|---|
| | 250 observations | | | 1000 observations | | |
| Variable importance | Complete | MCAR | MAR | Complete | MCAR | MAR |
| **% TP in top 10** | 2.3 | 0.3 | 0.3 | 22.0 | 1.7 | 1.7 |
| **% TP in top 10 but 1** | 37.0 | 16.7 | 11.0 | 53.3 | 73.7 | 59.0 |

Table 2.24: **Accuracy** simulation study results for **MissForest-Random Forest** analysis with Harrell bootstrap validation: scenarios **S1** (without missing data, no assumption of moderation) and **S2** (with missing data, complete outcome, no assumption of moderation) based on 300 data sets of **20 variables** each (n=250,1000). Means of all estimates along with their corresponding 2.5th and 97.5th percentile values within parenthesis. The theoretical MSE is $1.74^2 = 3.028$

| **Estimates** | 250 observations | | | 1000 observations | | |
|---|---|---|---|---|---|---|
| | **Complete cases** | **MCAR** | **MAR** | **Complete cases** | **MCAR** | **MAR** |
| $MSE_{apparent}$ | 4.837 (4.039,5.700) | 4.767 (3.911,5.780) | 4.584 (3.740,5.519) | 4.160 (3.816,4.505) | 3.924 (3.520,4.357) | 3.728 (3.267,4.184) |
| $\beta_{LP}$ | 1.140 (4.039,5.700) | 1.122 (3.911,5.780) | 1.124 (3.740,5.519) | 1.119 (3.816,4.505) | 1.106 (3.520,4.357) | 1.103 (3.267,4.184) |
| $MSE_{ext}$ | 5.087 (4.701,5.603) | 5.308 (4.744,5.951) | 5.087 (4.925,6.727) | 4.327 (4.093,4.530) | 4.549 (4.252,4.860) | 4.327 (4.369,5.313) |
| $Optimism_{ext}$ | -0.250 (-1.325,0.823) | -0.541 (-1.899,0.767) | -1.000 (-2.470,0.295) | -0.167 (-0.613,0.245) | -0.624 (-1.247,-0.053) | -1.022 (-1.889,-0.319) |
| $Optimism_{int}$ | -0.177 (-0.335,-0.060) | -0.492 (-0.721,-0.331) | -0.526 (-0.765,-0.353) | -0.165 (-0.231,-0.105) | -0.466 (-0.596,-0.366) | -0.483 (-0.600,-0.374) |
| $MSE_{corrected}$ | 5.014 (4.194,5.915) | 5.259 (4.299,6.483) | 5.111 (4.186,6.147) | 4.325 (3.957,4.674) | 4.390 (3.939,4.865) | 4.211 (3.711,4.706) |
| $\beta_{LP*}$ | 1.172 (1.129,1.220) | 1.181 (1.134,1.237) | 1.180 (1.130,1.240) | 1.145 (1.119,1.172) | 1.146 (1.117,1.179) | 1.145 (1.117,1.172) |

Table 2.25: **Accuracy** simulation study results for **MissForest-Random Forest** analysis with Harrell bootstrap validation: scenarios **S3** (assumption of moderation, without missing data) and **S4** (assumption of moderation, with missing data, complete outcome) based on 300 data sets of **20 variables** each (n=250,1000). Means of all estimates along with their corresponding 2.5th and 97.5th percentile values within parenthesis. The theoretical MSE is $1.74^2 = 3.028$

| **Estimates** | 250 observations | | | 1000 observations | | |
|---|---|---|---|---|---|---|
| | **Complete cases** | **MCAR** | **MAR** | **Complete cases** | **MCAR** | **MAR** |
| $MSE_{apparent}$ | 6.115 (5.154,7.148) | 6.079 (4.976,7.357) | 5.892 (4.836,7.047) | 5.071 (4.692,5.415) | 4.924 (4.464,5.439) | 4.666 (4.133,5.173) |
| $\beta_{LP}$ | 1.159 (5.154,7.148) | 1.142 (4.976,7.357) | 1.145 (4.836,7.047) | 1.134 (4.692,5.415) | 1.123 (4.464,5.439) | 1.123 (4.133,5.173) |
| $MSE_{ext}$ | 6.649 (6.061,7.342) | 6.867 (6.224,7.734) | 6.649 (6.271,8.086) | 5.400 (5.116,5.682) | 5.603 (5.257,6.060) | 5.400 (5.359,6.743) |
| $Optimism_{ext}$ | -0.534 (-1.933,0.723) | -0.788 (-2.384,0.524) | -1.153 (-2.772,0.330) | -0.330 (-0.855,0.156) | -0.679 (-1.430,-0.008) | -1.229 (-2.477,-0.322) |
| $Optimism_{int}$ | -0.118 (-0.245,0.007) | -0.463 (-0.688,-0.267) | -0.540 (-0.813,-0.309) | -0.095 (-0.145,-0.047) | -0.427 (-0.575,-0.329) | -0.516 (-0.709,-0.372) |
| $MSE_{corrected}$ | 6.292 (5.298,7.341) | 6.542 (5.376,7.831) | 6.432 (5.299,7.720) | 5.165 (4.799,5.527) | 5.351 (4.871,5.926) | 5.182 (4.550,5.717) |
| $\beta_{LP*}$ | 1.153 (1.117,1.193) | 1.165 (1.126,1.209) | 1.168 (1.127,1.220) | 1.126 (1.106,1.145) | 1.134 (1.112,1.156) | 1.136 (1.108,1.164) |

**MissForest-Random Forests S2: Missing data, No assumption of moderation, complete outcome** MissForest combined with RF performance in presence of missing data was very similar to the 15% tolerance model performance of MICE-Lasso and MissForest-Lasso. The optimism-corrected MSE was very poor: 5.259 (2.5th and 97.5th percentiles: 4.299 and 6.483) for MCAR 250 observation data (see Table 2.24 for MAR data). Estimates of optimism were smaller in absolute value than MICE-combined methods estimates and close to MissForest-Lasso estimates. Internal and external MSE optimism were very close for MCAR data, however there was bias in the estimate of internal optimism for MAR data.

MissForest-RF ranked the TPs among the top 10 variables with much less probability than the other methods. However, this probability was still higher than the probability of including FP, and the included FPs were mainly continuous FPs. Random Forest often tended to give more importance to the continuous noise variables with large probability of missingness (V13, see Figures 2.29 and 2.28).

**MissForest-Random Forests S3: No missing data, Assumption of moderation** After including interaction terms, the accuracy performance seemed to worsen as it happened with the other methods: average $MSE_{corrected}$=6.292 (2.5th and 97.5th percentiles: 5.298 and 7.341), but the average pseudo-$R^2$ was similarly 0.583 (2.5th and 97.5th percentiles: 0.511 and 0.649) for the 250 observations dataset. Again the estimates of optimism were the smallest among all the methods in absolute value. However, there was some discrepancy between internal and external MSE optimism (the latter being the largest), meaning that the internal validation process was giving too optimistic results compared to the external validation (see Table 2.25 and figure 2.15).

Variable importance was again biased because of the different scale of the variables and results were worse than scenario S1 (see Figure 2.30 and Table 2.26): RF never had all the TPs in the top 10 most important variables in the 250 observation datasets. The results worsening when interaction variables were added to the linear predictor was consistent with the other methods performance worsening in this scenario. When the sample size was small relative to the number of covariates, and most covariates were noise variables, the methods accuracy diminished. In fact, the larger sample size analysis (n=1000) returned slightly better results (see tables 2.25, 2.26 and figures A.27).

**MissForest-Random Forests S4: Missing data, Assumption of moderation, complete outcome** In presence of missing data, the optimism increased in absolute value compared to

scenario S3 with complete data, showing the bias caused by missing data imputation uncertainty. As a consequence, the optimism-corrected MSE increased was very poor: 6.542 (2.5th and 97.5th percentiles being 5.376 and 7.831) for MCAR data and 6.432 (2.5th and 97.5th percentiles being 5.299 and 7.720) for MAR data (see Table 2.25). The difference between internal and external MSE optimism was larger for MAR data as it happened in scenario S2.

The inclusion frequency of the TP in the top 10 most important variables was similar to the other MissForest-RF scenarios, i.e. continuous FP with a large percentage of missing data (V13) were considered as important as binary TPs (see Figures 2.31 and 2.33). Furthermore, the TPs were never exactly the top 10 variables in terms of importance (see Table 2.26).

**MissForest-Random Forests S5: Missing data also in outcome (20% missingness MAR and MCAR), Assumption of moderation** MissForest-RF had better discrimination performance when the outcome was not complete: $\text{MSE}_{corrected}$=5.632 (2.5th and 97.5th percentiles being 4.558 and 6.744) for MCAR data and 5.469 (2.5th and 97.5th percentiles being 4.301 and 6.613) in the case of 250 observations (see Table 2.28). As a consequence, the difference between internal and external MSE optimism was the largest among the other methods (see Figure 2.16) with the internal optimism being the least in absolute value and the external being the largest compared to the other method estimates.

Variable importance performance was always too parsimonious with respect to the binary TPs: their inclusion frequency ranged between 7% and 60% (see Figures 2.29 and 2.32). Again MissForest-RF never ranked all the TPs as most important variables at the same time (see Table 2.27)

Table 2.26: **Variable selection** simulation study results for scenarios **S3** (assumption of moderation, without missing data) and **S4** (assumption of moderation, with missing data) for **Random Forets** and **MissForest-Random Forests** best models in the case of **20 covariates** and 300 samples of 250 and 1000 observations. The following results are shown: the percentages of the times the TP are the top ranked variables among the selected, the percentage of times the TP are the top ranked variables apart from one TP.

| | **MissForest-RANDOM FOREST** | | | | | |
|---|---|---|---|---|---|---|
| | 250 observations | | | 1000 observations | | |
| Variable importance | Complete | MCAR | MAR | Complete | MCAR | MAR |
| **% TP in top 10** | 0 | 0 | 0 | 0 | 0 | 0 |
| **% TP in top 10 but 1** | 0 | 0 | 0 | 0.3 | 0 | 0 |

Table 2.27: **Variable selection** simulation study results for scenarios **S3** (assumption of moderation, without missing data) and **S5** (assumption of moderation, with missing data also in the outcome) for **Random Forests** and **MissForest-Random Forests** (MR) best models in the case of **20 covariates** and 300 samples of 250 and 1000 observations. The following results are shown: the percentages of the times the TP are the top ranked variables among the selected, the percentage of times the TP are the top ranked variables apart from one TP.

| | **MissForest-RANDOM FOREST** | | | | | |
|---|---|---|---|---|---|---|
| | 250 observations | | | 1000 observations | | |
| Variable importance | Complete | MCAR | MAR | Complete | MCAR | MAR |
| **% TP in top 10** | 0 | 0 | 0 | 0 | 0 | 0 |
| **% TP in top 10 but 1** | 0 | 0 | 0 | 0.3 | 0 | 0 |

Table 2.28: **Accuracy** simulation study results for **MissForest-Random Forest** analysis with Harrell bootstrap validation: scenarios **S3** (assumption of moderation, complete data) and **S5** (assumption of moderation, missing data also in the outcome), based on 300 data sets of **20 variables** each (n=250,1000). Means of all estimates along with their corresponding 2.5th and 97.5th percentile values within parenthesis. The theoretical MSE is $1.74^2 = 3.028$

| **Estimates** | 250 observations | | | 1000 observations | | |
|---|---|---|---|---|---|---|
| | **Complete cases** | **MCAR** | **MAR** | **Complete cases** | **MCAR** | **MAR** |
| $\text{MSE}_{apparent}$ | 6.115 (5.154,7.148) | 5.069 (4.034,6.111) | 4.893 (3.809,5.861) | 5.071 (4.692,5.415) | 4.148 (3.675,4.645) | 3.938 (3.489,4.325) |
| $\beta_{LP}$ | 1.159 (5.154,7.148) | 1.124 (4.034,6.111) | 1.131 (3.809,5.861) | 1.134 (4.692,5.415) | 1.106 (3.675,4.645) | 1.108 (3.489,4.325) |
| $\text{MSE}_{ext}$ | 6.649 (6.061,7.342) | 7.358 (6.584,8.340) | 6.649 (6.271,8.086) | 5.400 (5.116,5.682) | 5.990 (5.554,6.546) | 6.211 (5.673,6.974) |
| $\text{Optimism}_{ext}$ | -0.534 (-1.933,0.723) | -2.288 (-3.832,-0.842) | -2.620 (-4.371,-1.163) | -0.330 (-0.855,0.156) | -1.843 (-1.909,-1.161) | -2.273 (-3.313,-1.487) |
| $\text{Optimism}_{int}$ | -0.118 (-0.245,0.007) | -0.562 (-0.792,-0.384) | -0.576 (-0.868,-0.374) | -0.095 (-0.145,-0.047) | -0.468 (-0.578,-0.375) | -0.472 (-0.613,-0.359) |
| $\text{MSE}_{corrected}$ | 6.292 (5.298,7.341) | 5.632 (4.558,6.744) | 5.469 (4.301,6.613) | 5.165 (4.799,5.527) | 4.616 (4.107,5.146) | 4.409 (3.889,4.856) |
| $\beta_{LP^*}$ | 1.153 (1.117,1.193) | 1.151 (1.114,1.199) | 1.149 (1.104,1.201) | 1.126 (1.106,1.145) | 1.122 (1.100,1.144) | 1.118 (1.092,1.144) |

Figure 2.8: **Optimism-corrected MSE** estimates from 4 methods run on 300 simulated **20-covariate** datasets with **250 observations** for scenarios S1 (without missing data) and S2 (with missing data). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME), MissForest-Lasso (MF) and MissForest-Random Forests (MR). ML, ME and MF estimated MSEs are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S1 (first plot from the left), the Lasso (L), Elasticnet (E) and Random Forest (RF) corrected MSEs are shown.

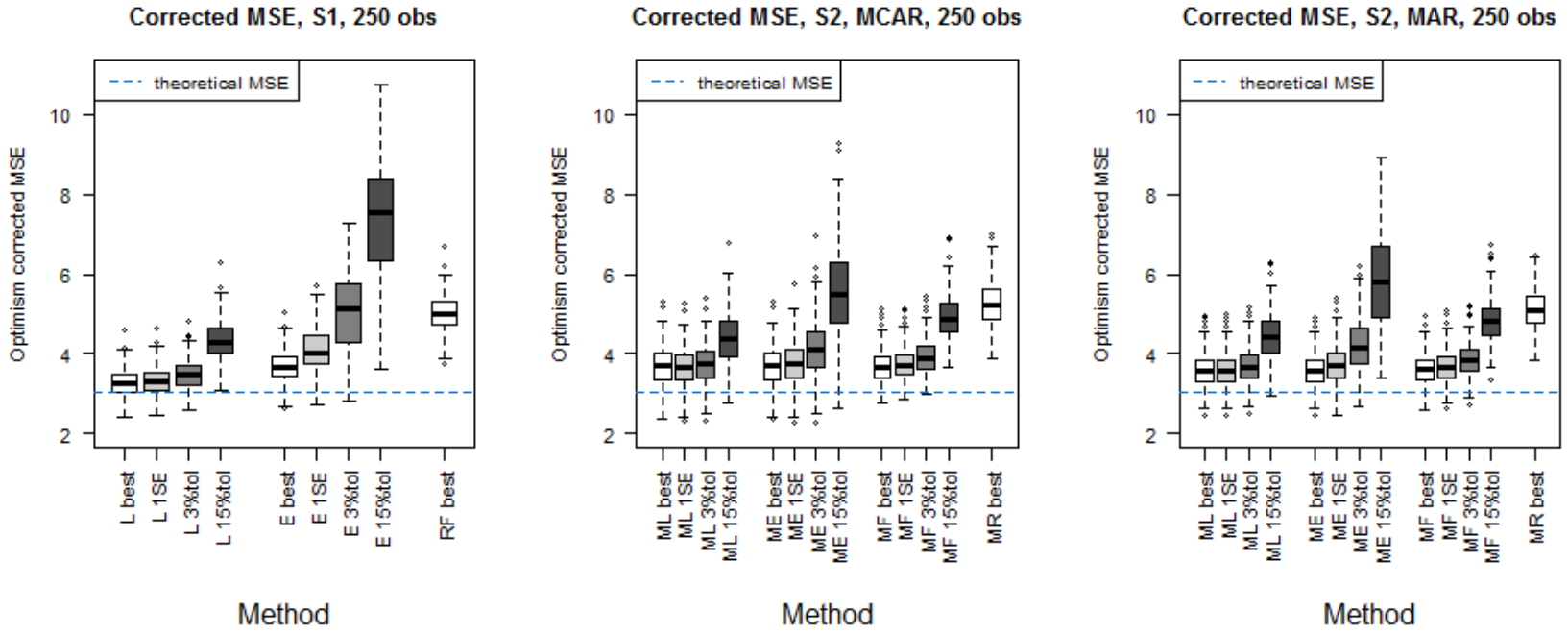


Figure 2.9: **Optimism-corrected MSE** estimates from 4 methods run on 300 simulated **20-covariate** datasets with **250 observations** for scenarios **S3** (assumption of moderation, without missing data) and **S4** (assumption of moderation, with missing data). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME), MissForest-Lasso (MF) and MissForest-Random Forests (MR). ML, ME and MF estimated MSEs are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S3 (first plot from the left), the Lasso (L), Elasticnet (E) and Random Forest (RF) corrected MSEs are shown.

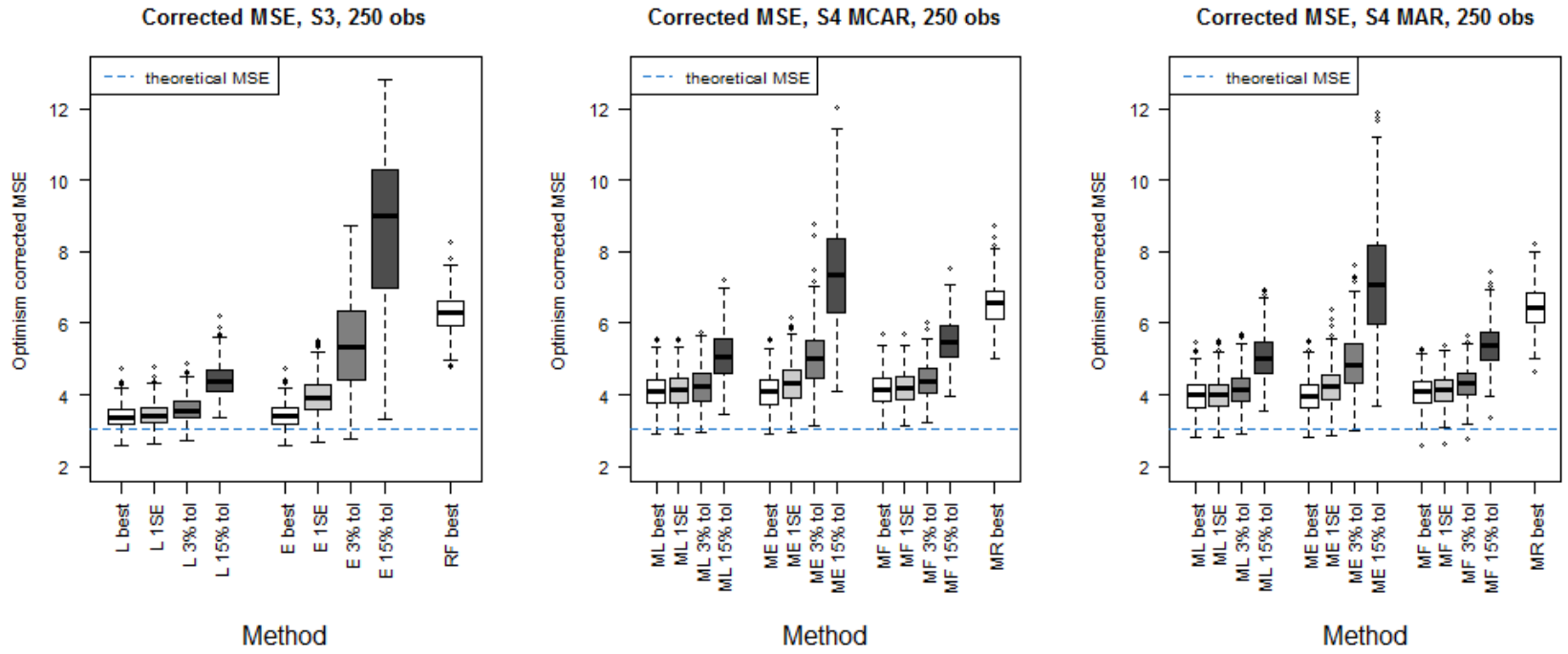


Figure 2.10: **Optimism-corrected MSE** estimates from 4 methods run on 300 simulated **20-covariate** datasets with 250 observations for scenarios **S3** (assumption of moderation, without missing data) and **S5** (assumption of moderation, with missing data also in the outcome). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME), MissForest-Lasso (MF) and MissForest-Random Forests (MR). ML, ME and MF estimated MSEs are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S3 (first plot from the left), the Lasso (L), Elasticnet (E) and Random Forest (RF) corrected MSEs are shown.

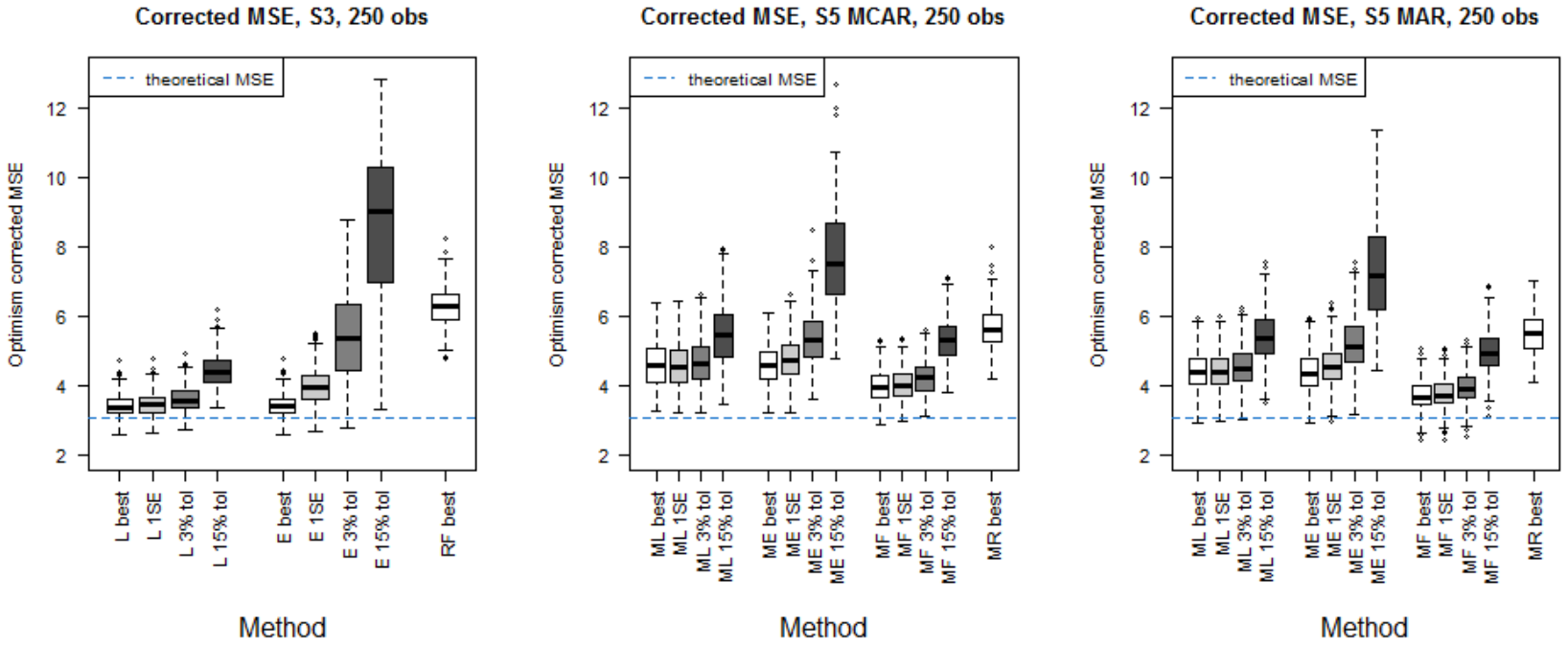


Figure 2.11: **Calibration slope** $\beta_{LP}$ estimates for 4 methods run on 300 simulated **20-covariate** datasets with **250 observations** for scenarios S1 (without missing data) and S2 (with missing data). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME), MissForest-Lasso (MF) and MissForest-Random Forests (MR). ML, ME and MF estimated calibration slopes are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S1 (first plot from the left), the Lasso (L), Elasticnet (E) and Random Forest (RF) calibration slopes are shown.

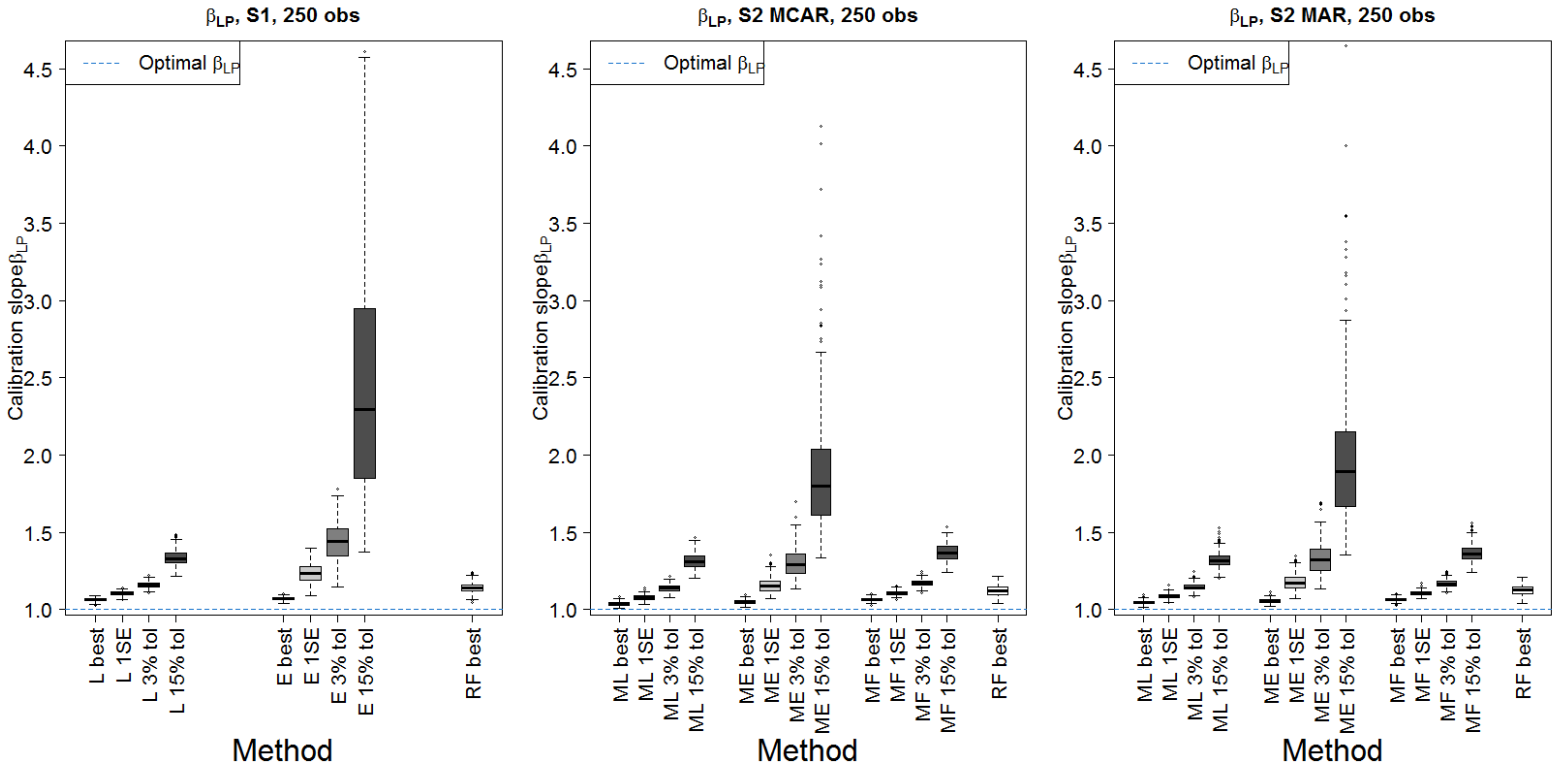


Figure 2.12: **Calibration slope** $\beta_{LP}$ estimates for 4 methods run on 300 simulated **20-covariate** datasets with **250 observations** for scenarios **S3** (assumption of moderation, without missing data) and **S4** (assumption of moderation, with missing data). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME), MissForest-Lasso (MF) and MissForest-Random Forests (MR). ML, ME and MF estimated calibration slopes are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S3 (first plot from the left), the Lasso (L), Elasticnet (E) and Random Forest (RF) calibration slopes are shown.

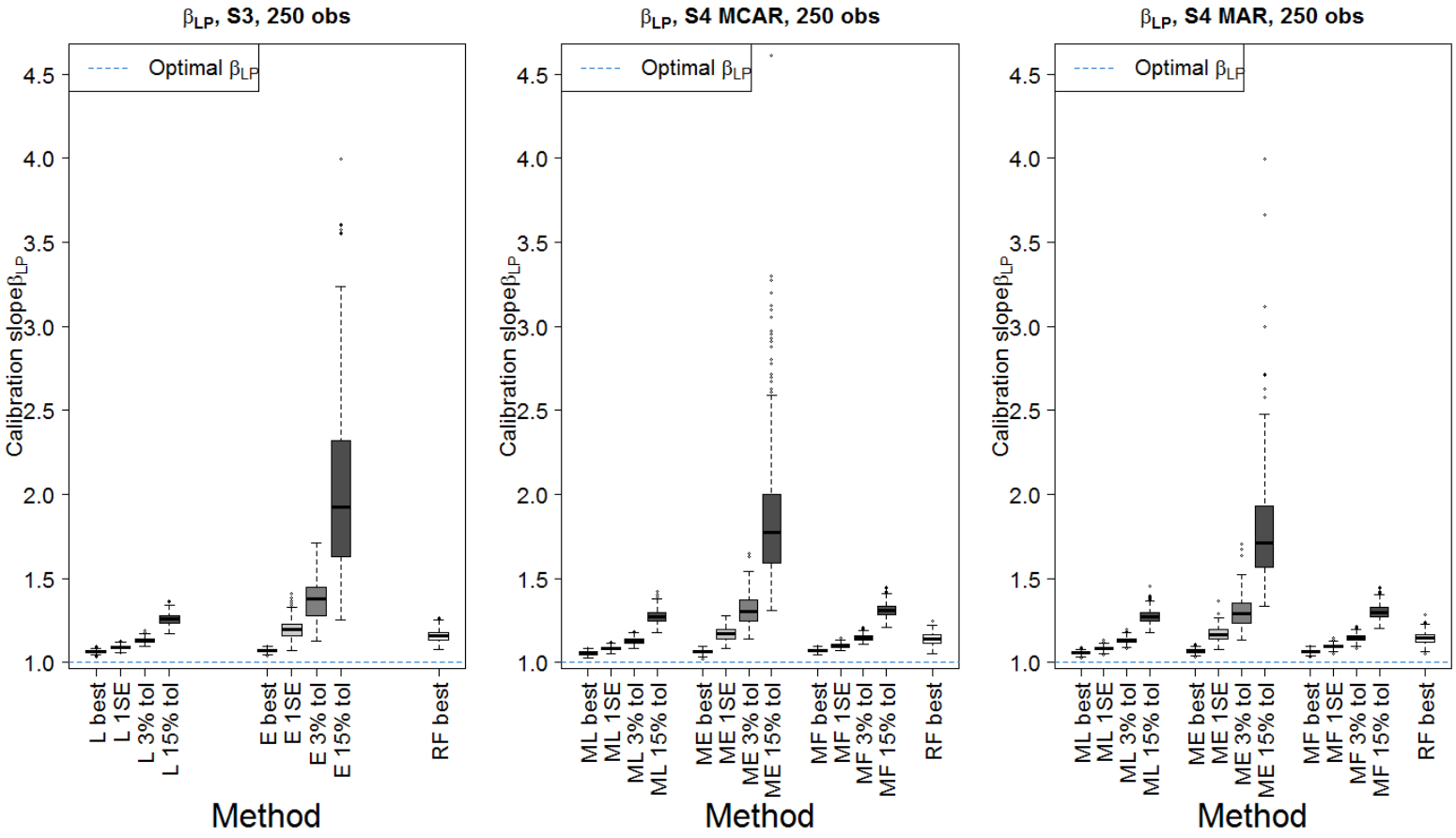


Figure 2.13: **Calibration slope** $\beta_{LP}$ estimates for 4 methods run on 300 simulated **20-covariate** datasets with **250 observations** for scenarios **S3** (assumption of moderation, without missing data) and **S5** (assumption of moderation, missing data also in the outcome). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME), MissForest-Lasso (MF) and MissForest-Random Forests (MR). ML, ME and MF estimated calibration slopes are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S3 (first plot from the left), the Lasso (L), Elasticnet (E) and Random Forest (RF) calibration slopes are shown.

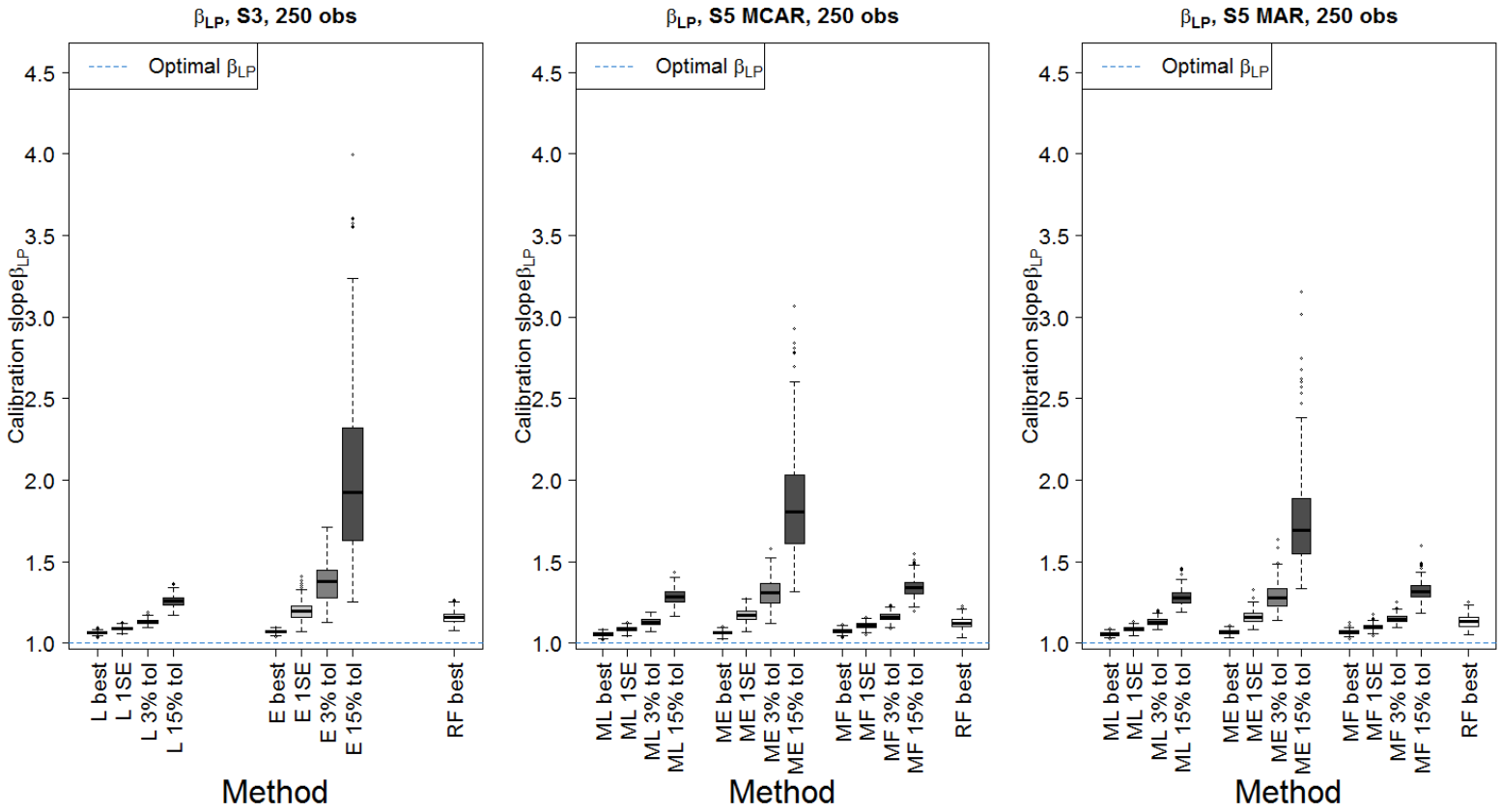


Figure 2.14: Average **internal and external MSE optimism** estimates with 2.5th and 97.5th percentiles for 4 methods run on 300 simulated **20-covariate** datasets with **250 observations** for scenarios **S1** (without missing data) and **S2** (with missing data). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME), MissForest-Lasso (MF) and MissForest-Random Forests (MR). ML, ME and MF estimated internal and external MSE optimism are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S1 (first plot from the left), the Lasso (L), Elasticnet (E) and Random Forest (RF) optimism estimates are shown.

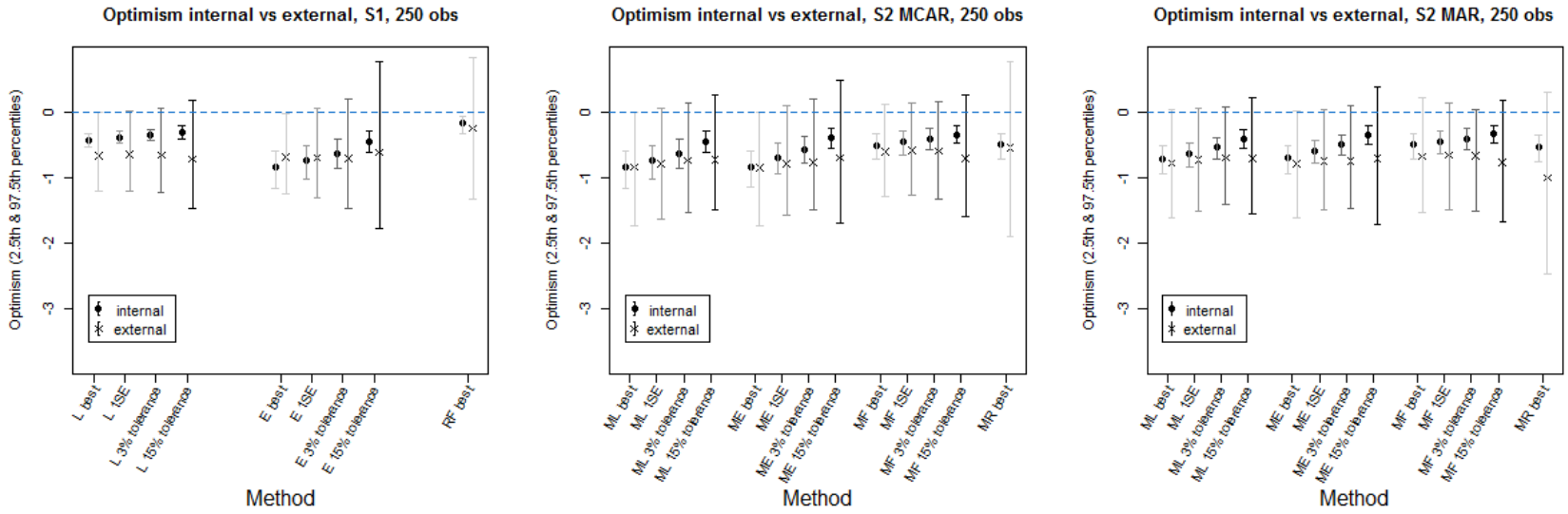


Figure 2.15: Average **internal and external MSE optimism** estimates with 2.5th and 97.5th percentiles for 4 methods run on 300 simulated **20-covariate** datasets with **250 observations** for scenarios **S3** (assumption of moderation, without missing data) and **S4** (assumption of moderation, with missing data). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME), MissForest-Lasso (MF) and MissForest-Random Forests (MR). ML, ME and MF estimated internal and external MSE optimism are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S3 (first plot from the left), the Lasso (L), Elasticnet (E) and Random Forest (RF) optimism estimates are shown.

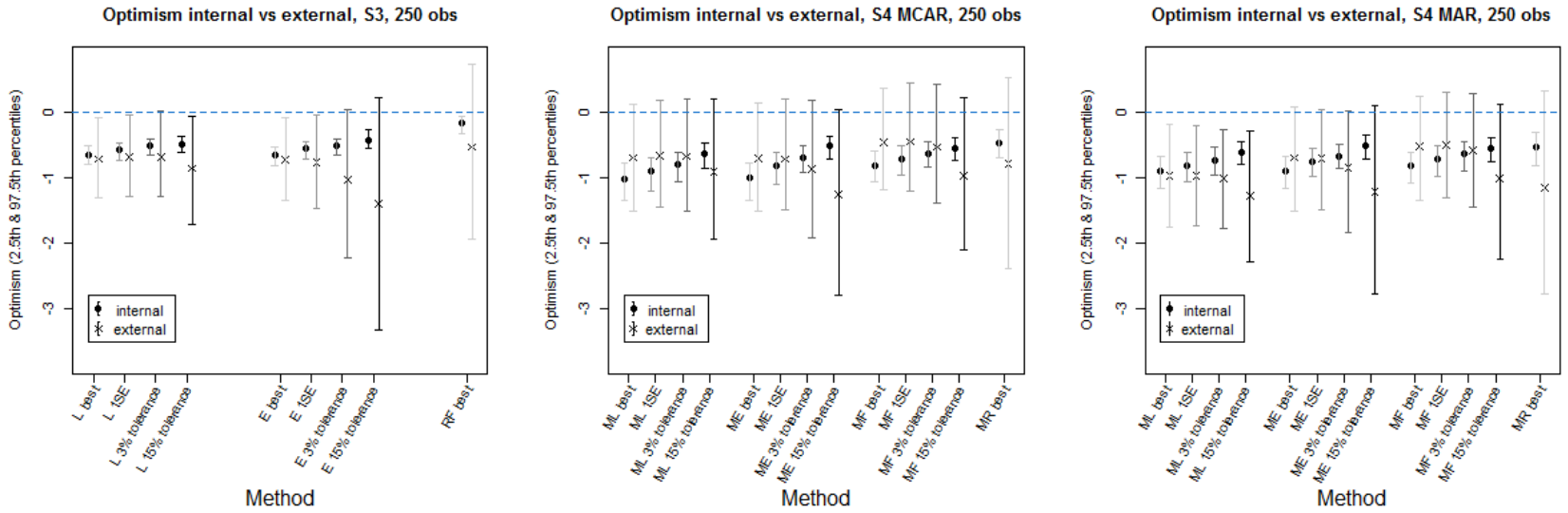


Figure 2.16: Average **internal and external MSE optimism** estimates with 2.5th and 97.5th percentiles for 4 methods run on 300 simulated **20-covariate** datasets with **250 observations** for scenarios **S3** (assumption of moderation, without missing data) and **S5** (assumption of moderation, with missing data also in the outcome). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME), MissForest-Lasso (MF) and MissForest-Random Forests (MR). ML, ME and MF estimated internal and external MSE optimism are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S3 (first plot from the left), the Lasso (L), Elasticnet (E) and Random Forest (RF) optimism estimates are shown.

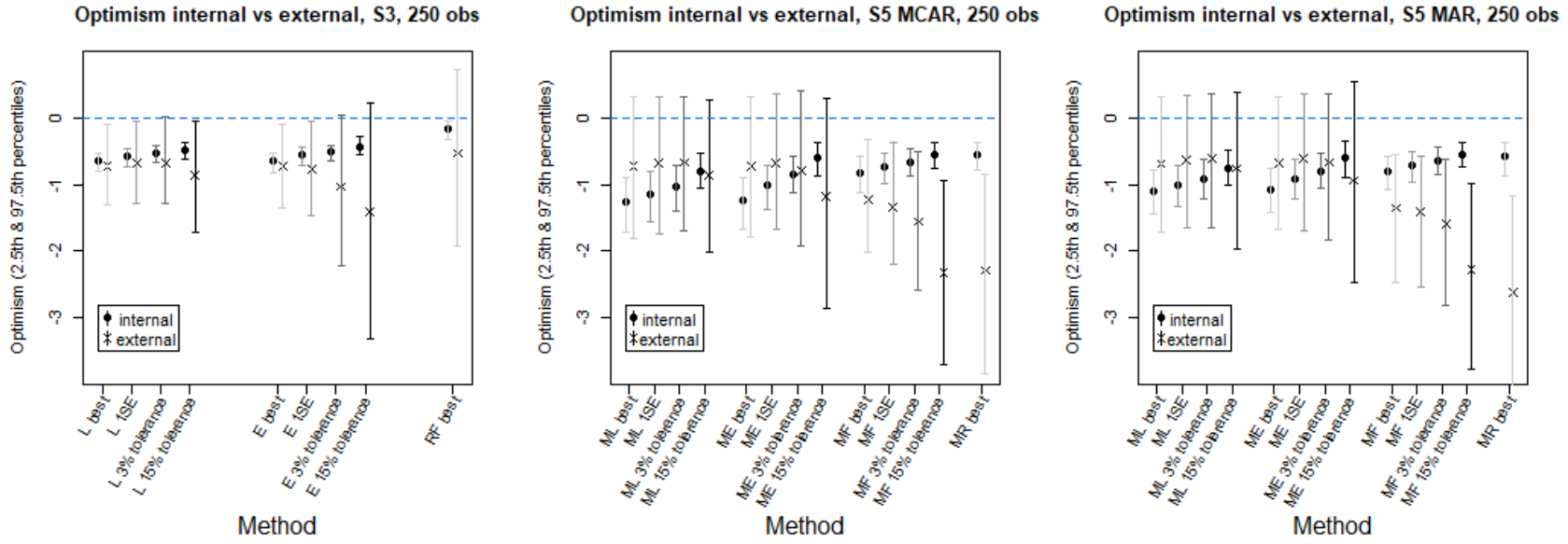


Figure 2.17: Average percentage of **true predictors (TP) selected among the actual TP** (SEN) estimates with 2.5th and 97.5th percentiles from 3 methods run on 300 simulated **20-covariate** datasets with **250 observations** for scenarios **S1** (without missing data) and **S2** (with missing data). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME) and MissForest-Lasso (MF). ML, ME and MF estimated percentages of TP selected among the actual TP variables are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S1 (first plot from the left), the Lasso (L) and the Elasticnet (E) estimates are shown.

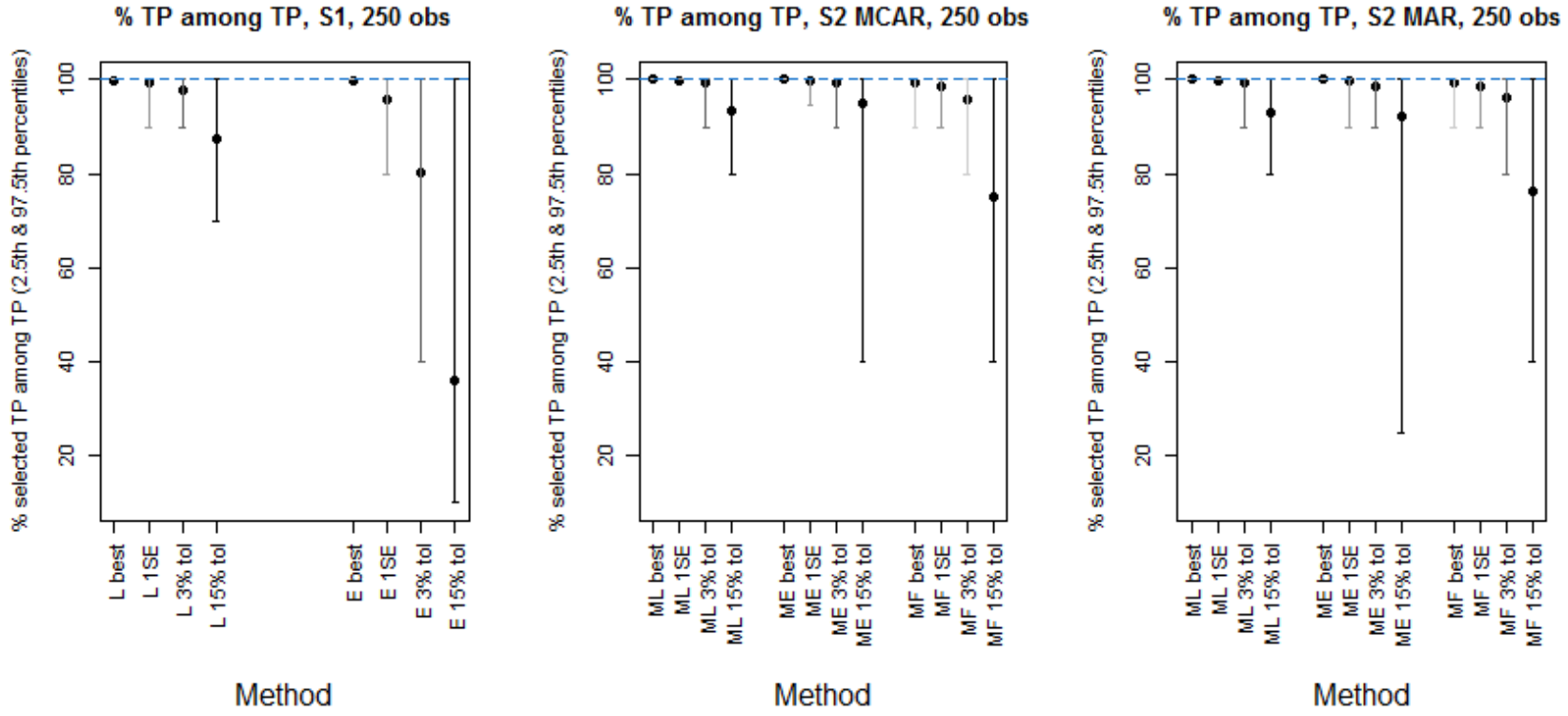


Figure 2.18: Average percentage of **true predictors (TP) selected among the actual TP** (SEN) estimates with 2.5th and 97.5th percentiles from 3 methods run on 300 simulated **20-covariate** datasets with **250 observations** for scenarios **S3** (assumption of moderation, without missing data) and **S4** (assumption of moderation, with missing data). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME) and MissForest-Lasso (MF). ML, ME and MF estimated percentages of TP selected among the actual TP variables are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S3 (first plot from the left), the Lasso (L) and the Elasticnet (E) estimates are shown.

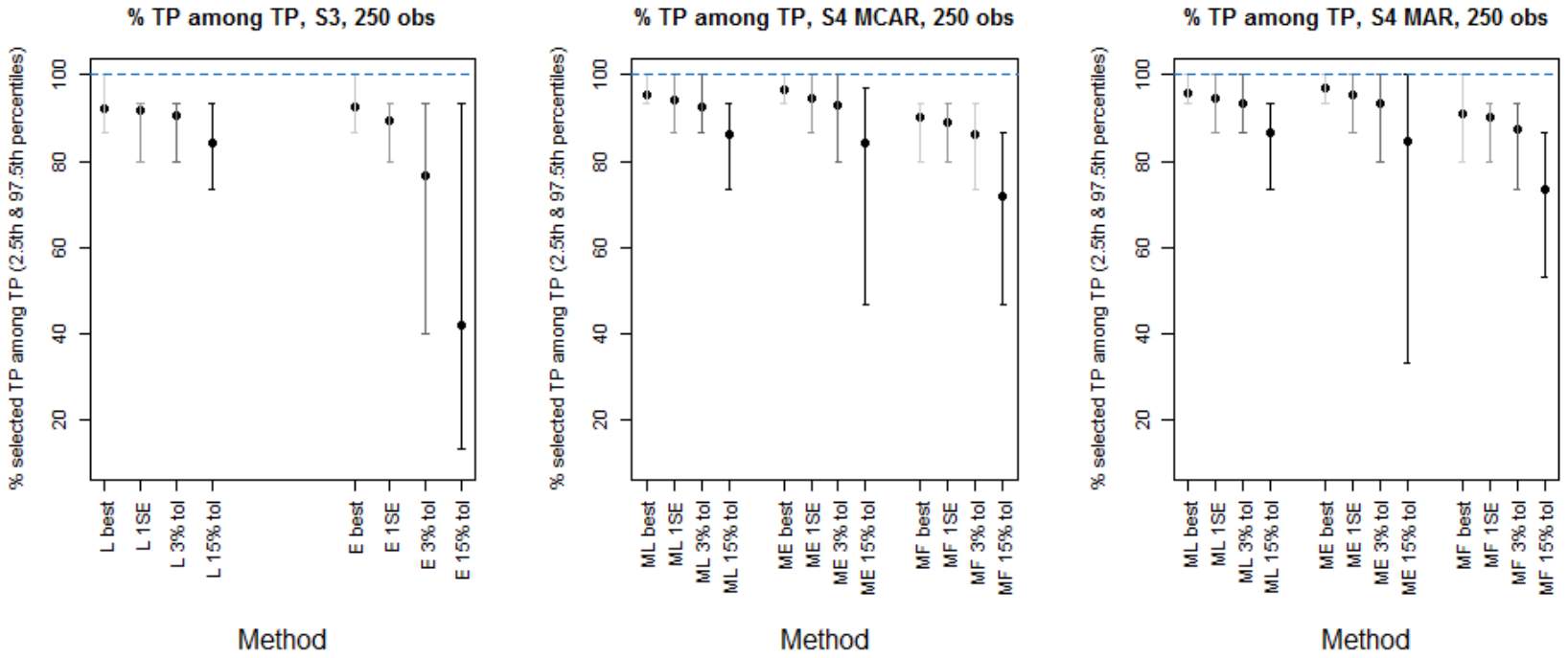


Figure 2.19: Average percentage of **true predictors (TP) selected among the actual TP** (SEN) estimates with 2.5th and 97.5th percentiles from 3 methods run on 300 simulated **20-covariate** datasets with **250 observations** for scenarios **S3** (assumption of moderation, without missing data) and **S5** (assumption of moderation, missing data also in the outcome). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME) and MissForest-Lasso (MF). ML, ME and MF estimated percentages of TP selected among the actual TP variables are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S3 (first plot from the left), the Lasso (L) and the Elasticnet (E) estimates are shown.

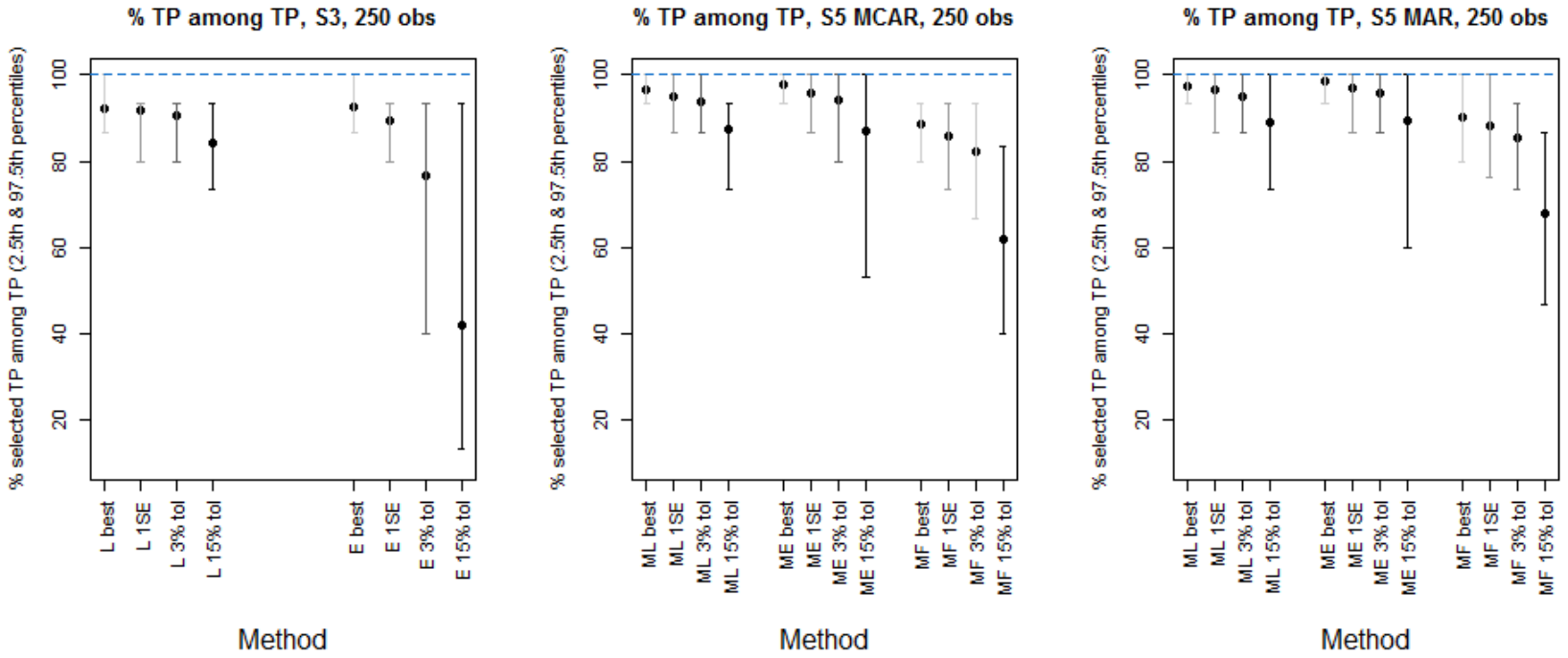


Figure 2.20: Average percentage of **true predictors (TP) among the selected variables** (PPV) estimates with 2.5th and 97.5th percentiles from 3 methods run on 300 simulated **20-covariate** datasets with **250 observation** for scenarios **S1** (without missing data) and **S2** (with missing data). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME) and MissForest-Lasso (MF). ML, ME and MF estimated percentages of TP among the selected variables are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S1 (first plot from the left), the Lasso (L) and the Elasticnet (E) estimates are shown.

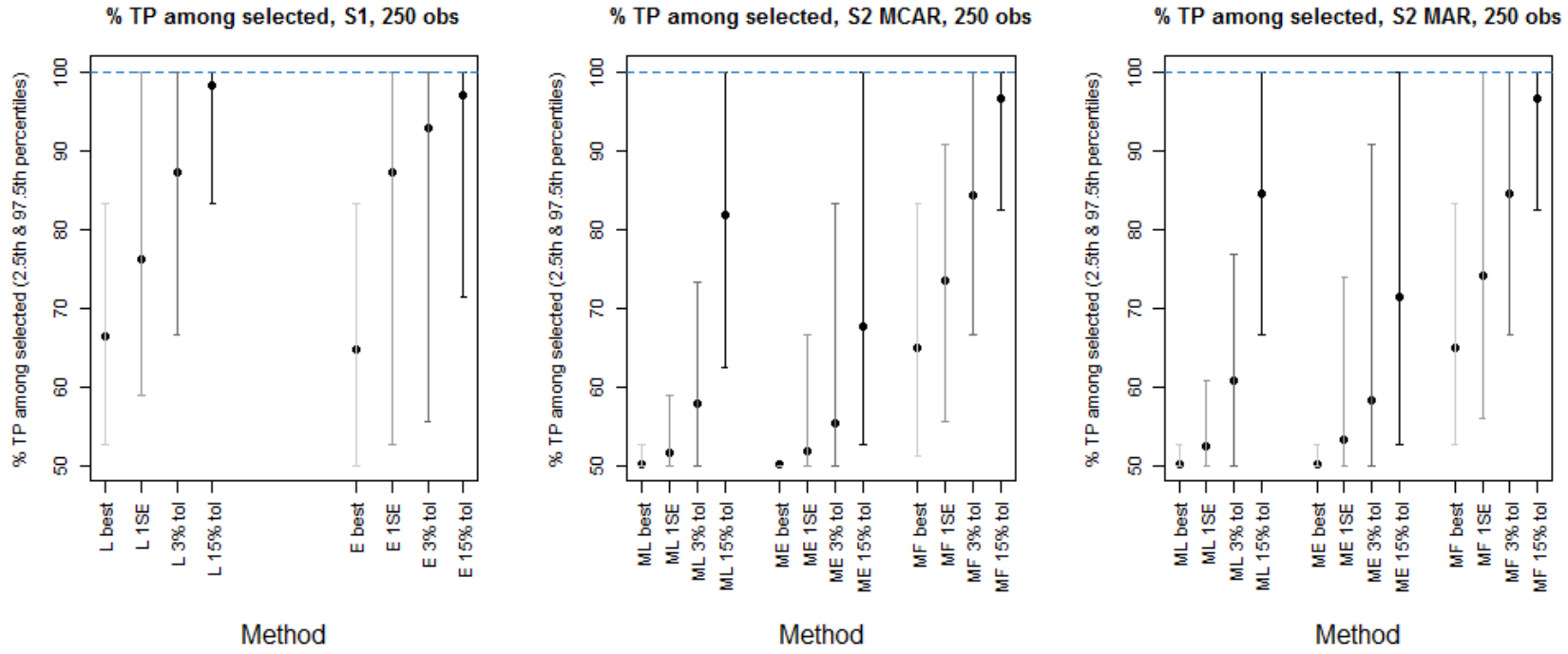


Figure 2.21: Average percentage of **true predictors (TP) among the selected variables** (PPV) estimates with 2.5th and 97.5th percentiles from 3 methods run on 300 simulated **20-covariate** datasets with **250 observations** for scenarios **S3** (assumption of moderation, without missing data) and **S4** (assumption of moderation, with missing data). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME) and MissForest-Lasso (MF). ML, ME and MF estimated percentages of TP among the selected variables are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S3 (first plot from the left), the Lasso (L) and the Elasticnet (E) estimates are shown.

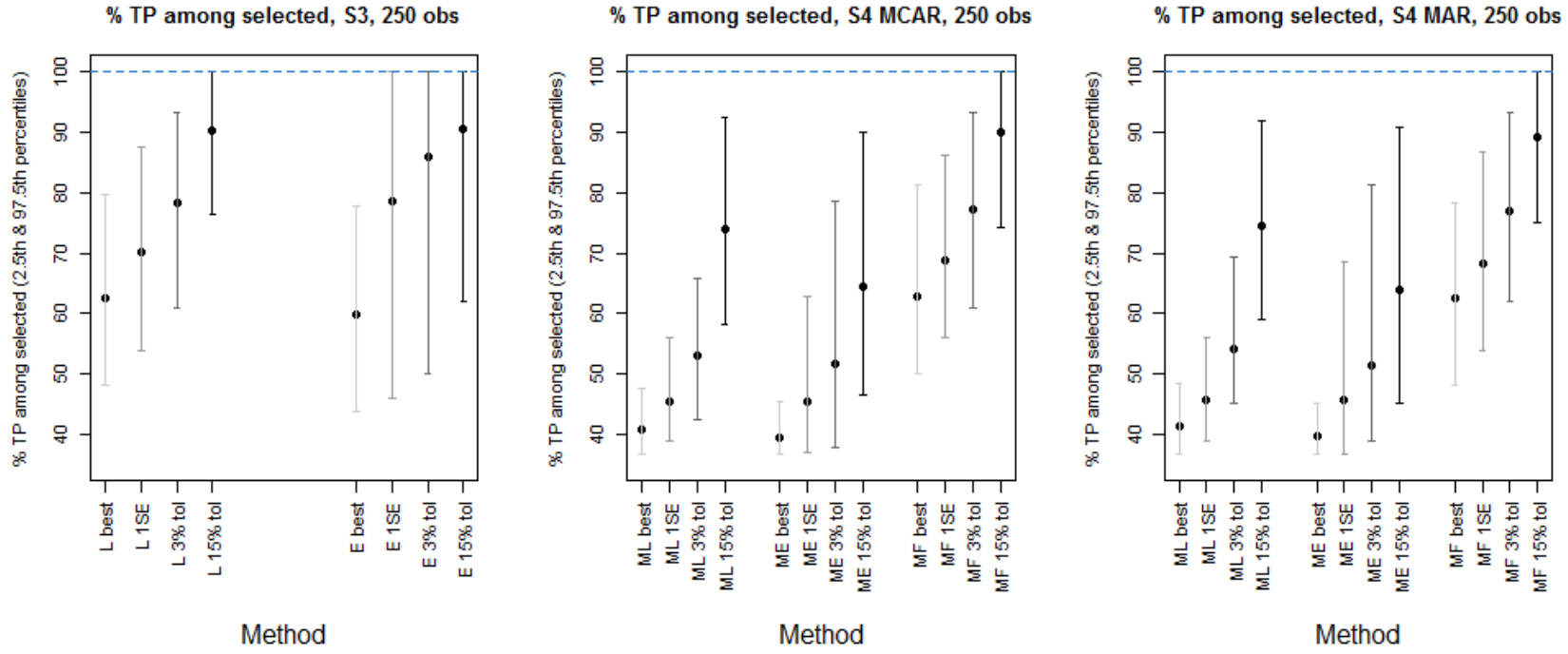


Figure 2.22: Average percentage of **true predictors (TP) among the selected variables** (PPV) estimates with 2.5th and 97.5th percentiles from 3 methods run on 300 simulated **20-covariate** datasets with **250 observations** for scenarios **S3** (assumption of moderation, without missing data) and **S5** (assumption of moderation, missing data also in the outcome). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME) and MissForest-Lasso (MF). ML, ME and MF estimated percentages of TP among the selected variables are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S3 (first plot from the left), the Lasso (L) and the Elasticnet (E) estimates are shown.

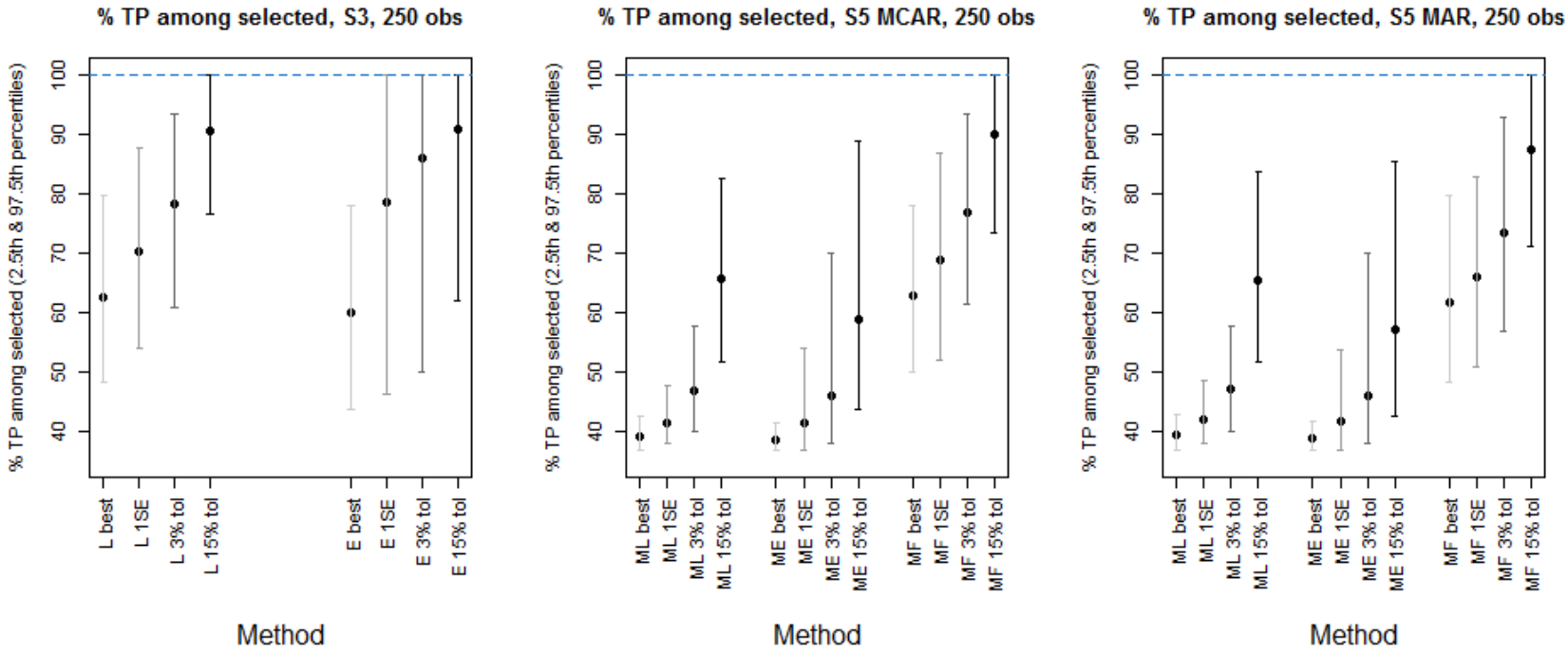


Figure 2.23: Estimated percentage of **correct (true) models** (simultaneously with respect to all predictors) found by 4 methods run on 300 simulated **20-covariate** datasets with **250 observations** for scenarios **S1** (without missing data) and **S2** (with missing data, complete outcome). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME), MissForest-Lasso (MF) and MissForest-Random Forests (MR). ML, ME and MF estimated percentages of selected true models are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S1 (first plot from the left), the Lasso (L), the Elasticnet (E) and the Random Forest (RF) estimates are shown. For the models RF and MR it is assumed that the true model is returned when the top 10 important variables are true predictors.

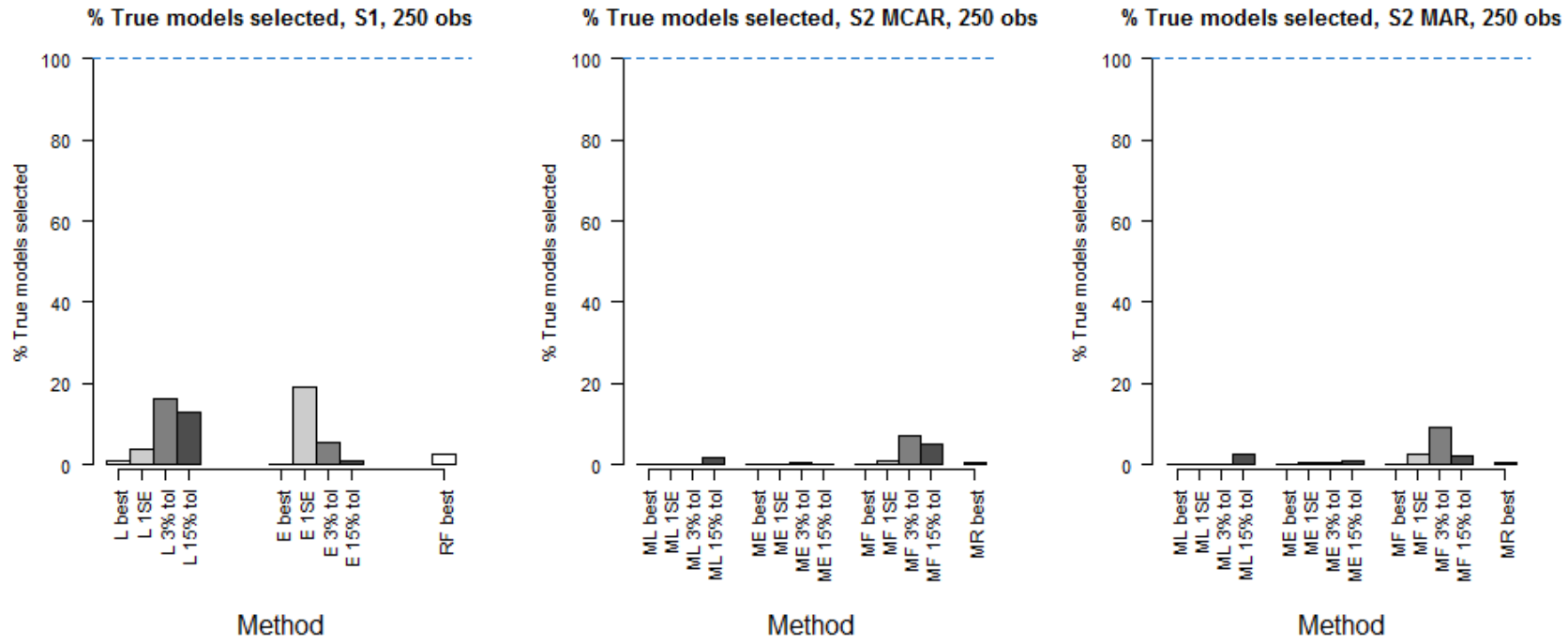


Figure 2.24: Estimated percentage of **almost correct models (only one variable off)** found by 4 methods run on 300 simulated **20-covariate** datasets with **250 observations** for scenarios **S1** (without missing data) and **S2** (with missing data, complete outcome). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME), MissForest-Lasso (MF) and MissForest-Random Forests (MR). ML, ME and MF estimated percentages of selected true models are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S1 (first plot from the left), the Lasso (L), the Elasticnet (E) and the Random Forests (RF) estimates are shown. For the models RF and MR, only the best model is computed through bootstrap tuning of the parameter given by the number of variables chosen randomly at each split to build the trees and it is assumed that the true model is returned when the top 10 important variables are the true predictors.

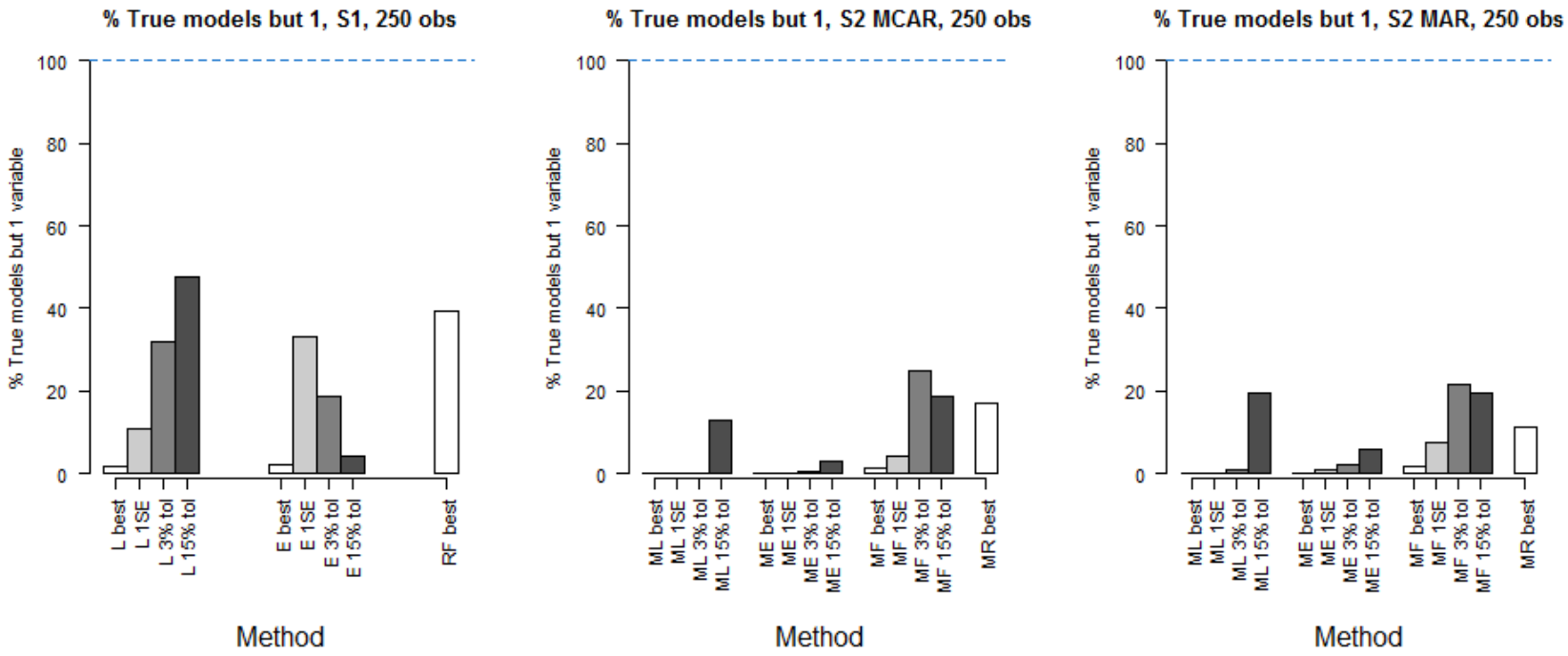


Figure 2.25: Estimated percentage of **almost correct models (only one variable off)** found by 4 methods run on 300 simulated **20-covariate** datasets with **250 observations** for scenarios **S3** (assumption of moderation, without missing data) and **S4** (assumption of moderation, with missing data). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME), MissForest-Lasso (MF) and MissForest-Random Forests (MR). ML, ME and MF estimated percentages of selected true models are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S3 (first plot from the left), the Lasso (L), the Elasticnet (E) and the Random Forests (RF) estimates are shown. For the models RF and MR, only the best model is computed through bootstrap tuning of the parameter given by the number of variables chosen randomly at each split to build the trees and it is assumed that the true model is returned when the top 10 important variables are the true predictors.

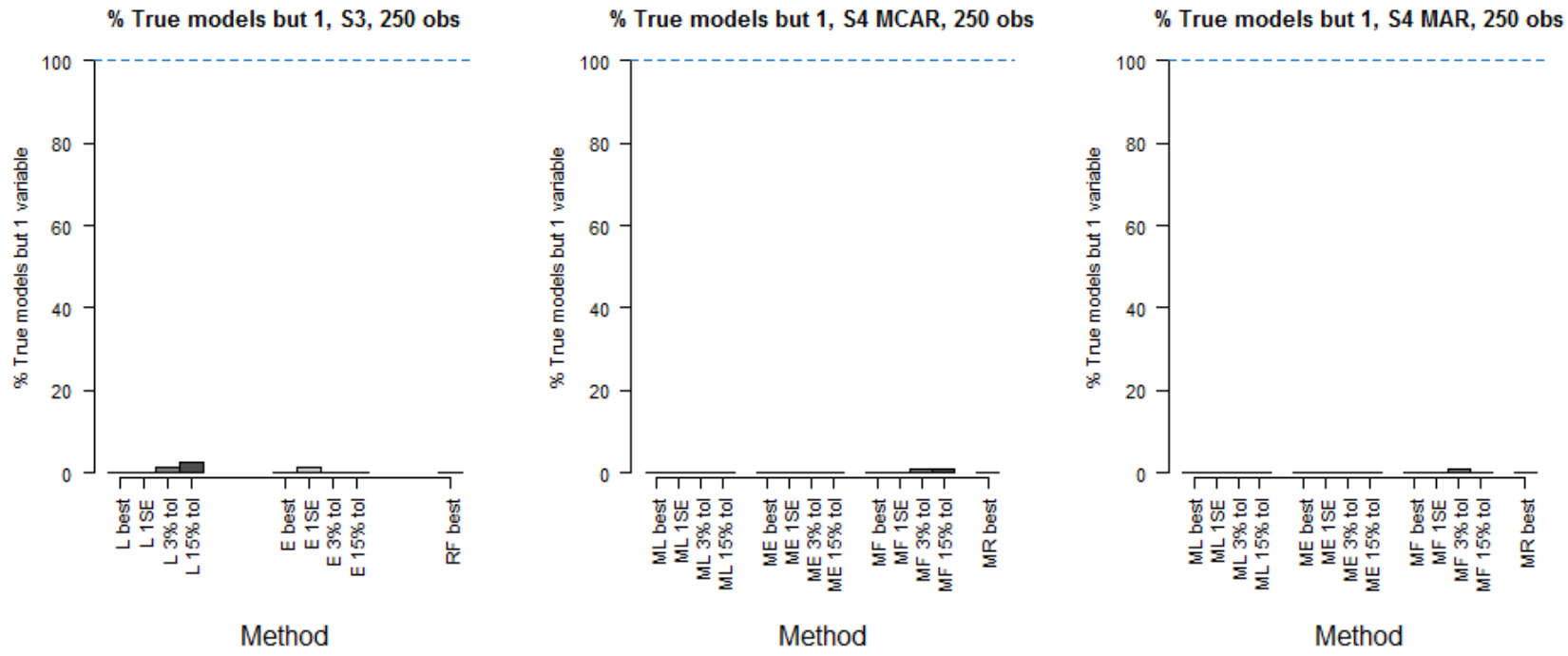


Figure 2.26: Estimated percentage of **almost correct models (only one variable off)** found by 4 methods run on 300 simulated **20-covariate** datasets with **250 observations** for scenarios **S3** (assumption of moderation, without missing data) and **S5** (assumption of moderation, missing data also in the outcome). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME), MissForest-Lasso (MF) and MissForest-Random Forests (MR). ML, ME and MF estimated percentages of selected true models are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S3 (first plot from the left), the Lasso (L), the Elasticnet (E) and the Random Forests (RF) estimates are shown. For the models RF and MR, only the best model is computed through bootstrap tuning of the parameter given by the number of variables chosen randomly at each split to build the trees and it is assumed that the true model is returned when the top 10 important variables are the true predictors.

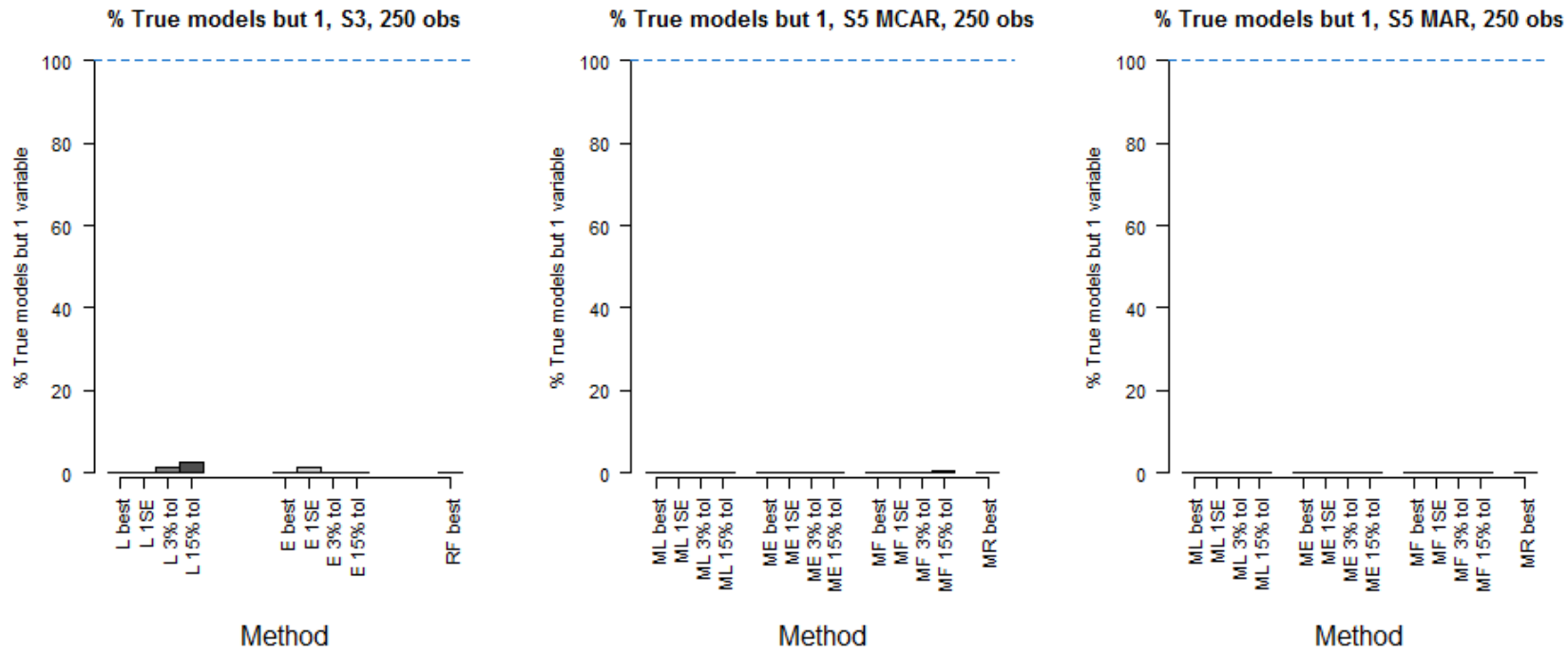


Figure 2.27: Comparison of **variable inclusion frequency** by 3 methods run on 300 simulated **20-covariate** datasets with **250 observations** for scenario **S1** (no assumption of moderation, complete data). The methods are: Lasso, Elasticnet and Random Forest. Lasso and Elasticnet variable inclusion frequencies are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For Random Forests only the best model is computed through bootstrap tuning of the parameter given by the number of variables chosen randomly at each split to build the trees and a variable is considered included in the model when its importance is among the top 10 variable importances.

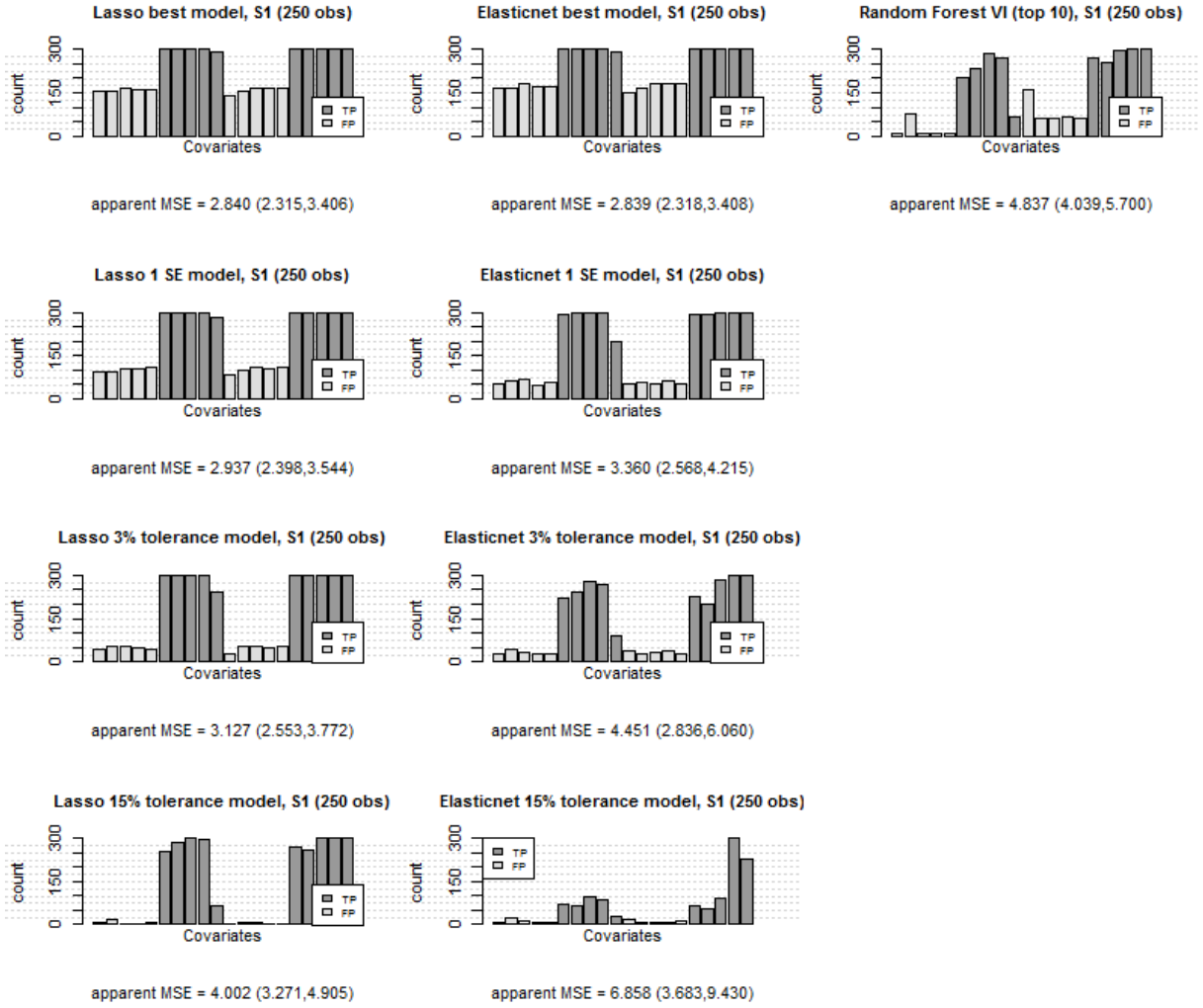


Figure 2.28: Comparison of **variable inclusion frequency** by 3 methods run on 300 simulated **20-covariate** datasets with **250 observations** for scenario **S2** with **MCAR** data (no assumption of moderation, complete outcome). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME), MissForest-Lasso (MF) and MissForest-Random Forests (MR). ML, ME and MF variable inclusion frequencies are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For MR only the best model is computed through bootstrap tuning of the parameter given by the number of variables chosen randomly at each split to build the trees and a variable is considered included in the model when its importance is among the top 10 variable importances.

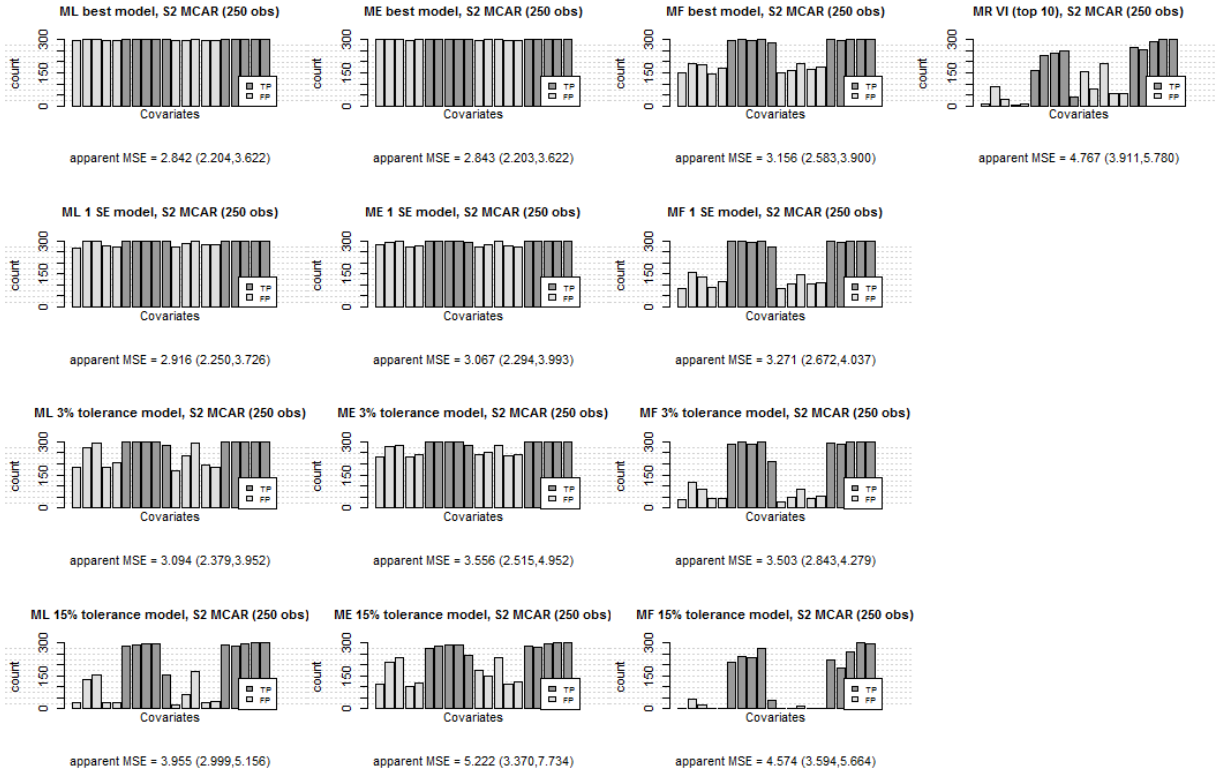


Figure 2.29: Comparison of **variable inclusion frequency** by 3 methods run on 300 simulated **20-covariate** datasets with **250 observations** for scenario **S2** with **MAR** data (no assumption of moderation, complete outcome). MICE-Lasso (ML), MICE-Elasticnet (ME), MissForest-Lasso (MF) and MissForest-Random Forests (MR). ML, ME and MF variable inclusion frequencies are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For MR only the best model is computed through bootstrap tuning of the parameter given by the number of variables chosen randomly at each split to build the trees and a variable is considered included in the model when its importance is among the top 10 variable importances.

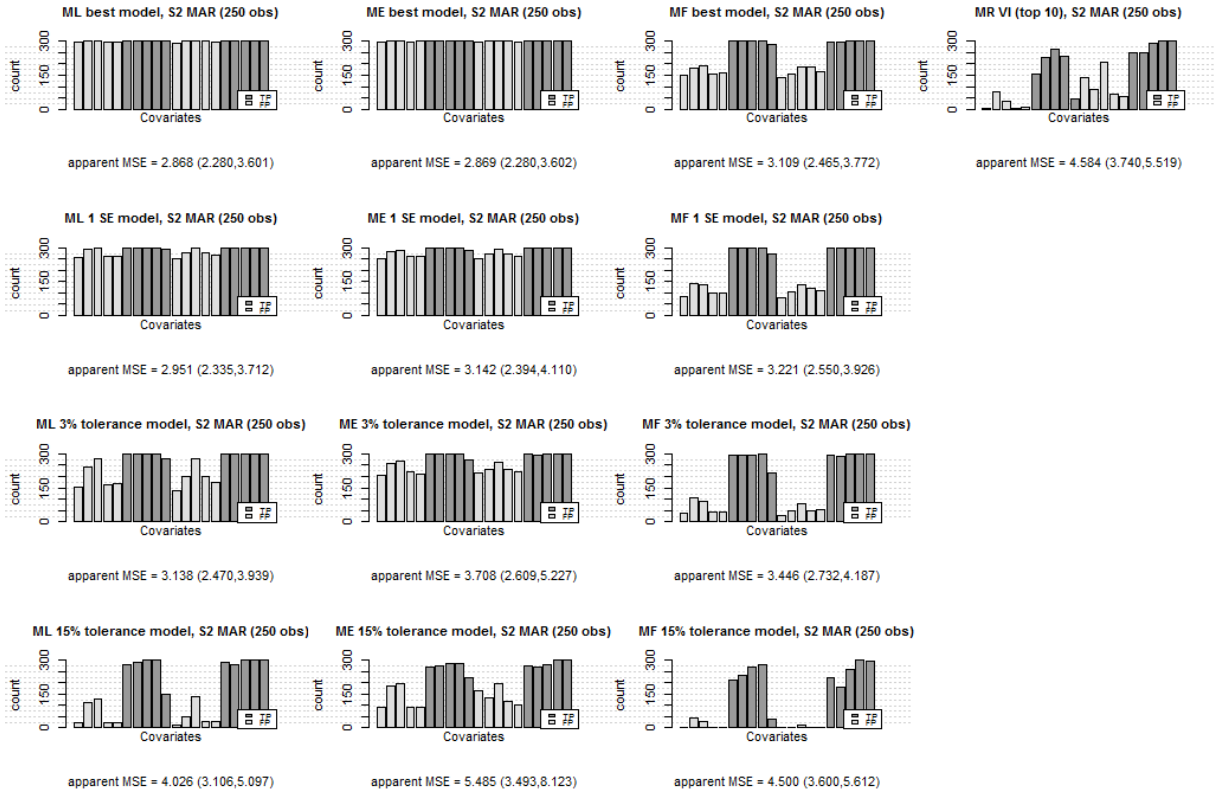


Figure 2.30: Comparison of **variable inclusion frequency** by 3 methods run on 300 simulated **20-covariate** datasets with **250 observations** for scenario **S3** (assumption of moderatio , complete data). The methods are: Lasso, Elasticnet and Random Forest. Lasso and Elasticnet variable inclusion frequencies are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For Random Forests only the best model is computed through bootstrap tuning of the parameter given by the number of variables chosen randomly at each split to build the trees and a variable is considered included in the model when its importance is among the top 10 variable importances.

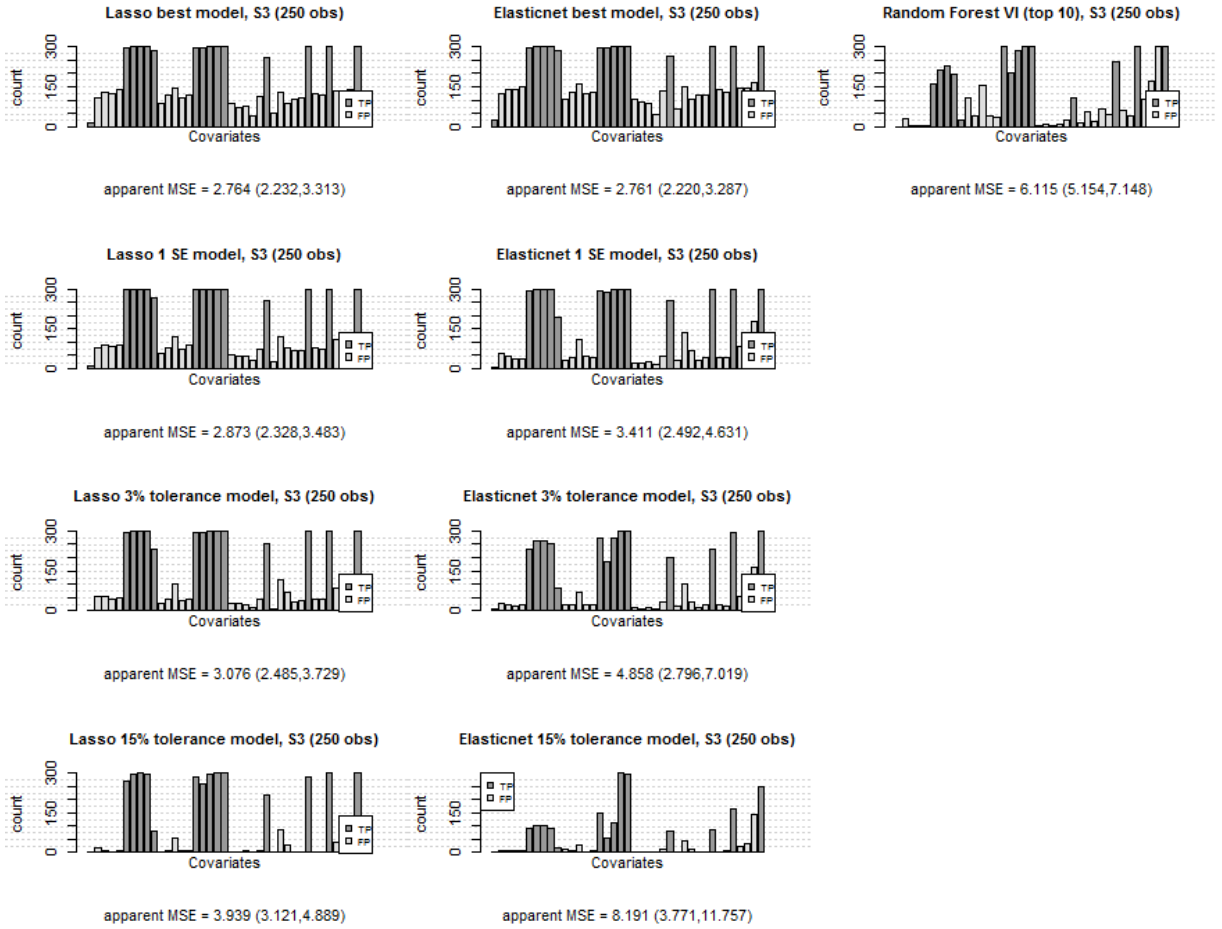


Figure 2.31: Comparison of **variable inclusion frequency** by 3 methods run on 300 simulated **20-covariate** datasets with **250 observations** for scenario **S4** with **MCAR** data (assumption of moderation, complete outcome). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME), MissForest-Lasso (MF) and MissForest-Random Forests (MR). ML, ME and MF variable inclusion frequencies are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For MR only the best model is computed through bootstrap tuning of the parameter given by the number of variables chosen randomly at each split to build the trees and a variable is considered included in the model when its importance is among the top 10 variable importances.

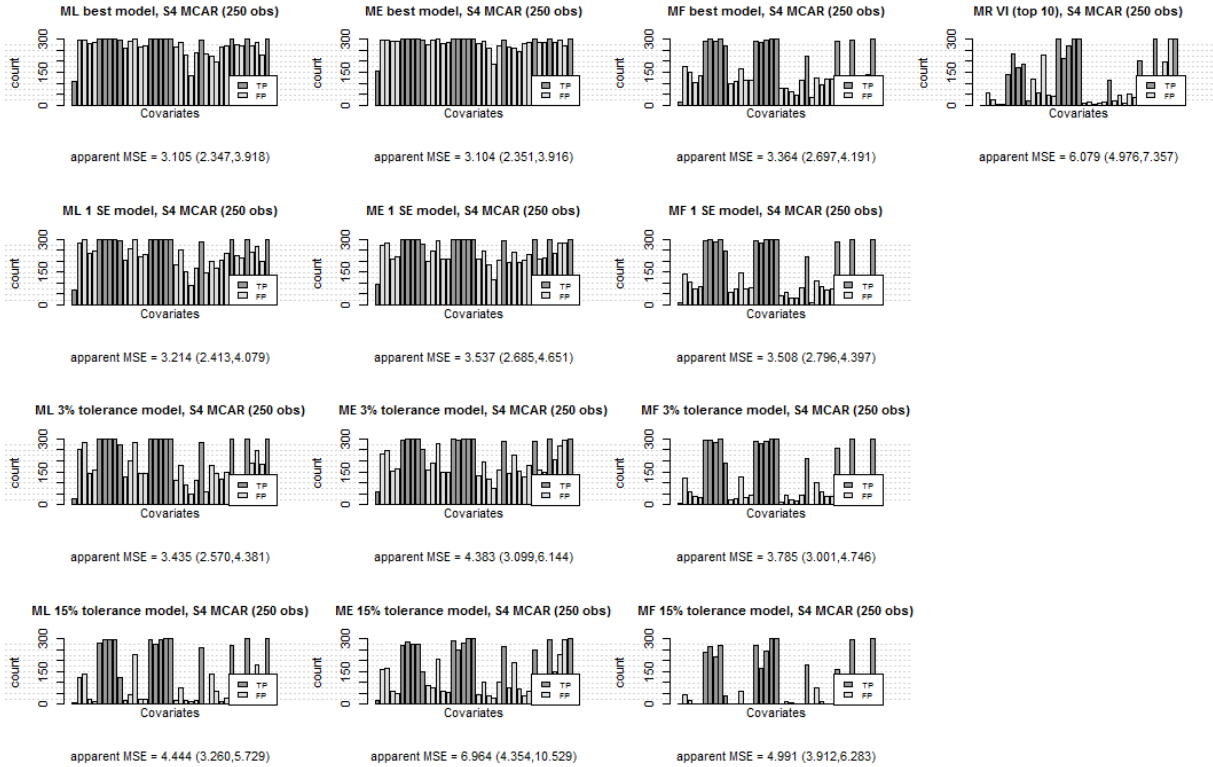


Figure 2.32: Comparison of **variable inclusion frequency** by 3 methods run on 300 simulated **20-covariate** datasets with 250 observations for scenario **S5** with **MCAR** data (assumption of moderation, missing data also in the outcome). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME), MissForest-Lasso (MF) and MissForest-Random Forests (MR). ML, ME and MF variable inclusion frequencies are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For MR only the best model is computed through bootstrap tuning of the parameter given by the number of variables chosen randomly at each split to build the trees and a variable is considered included in the model when its importance is among the top 10 variable importances.

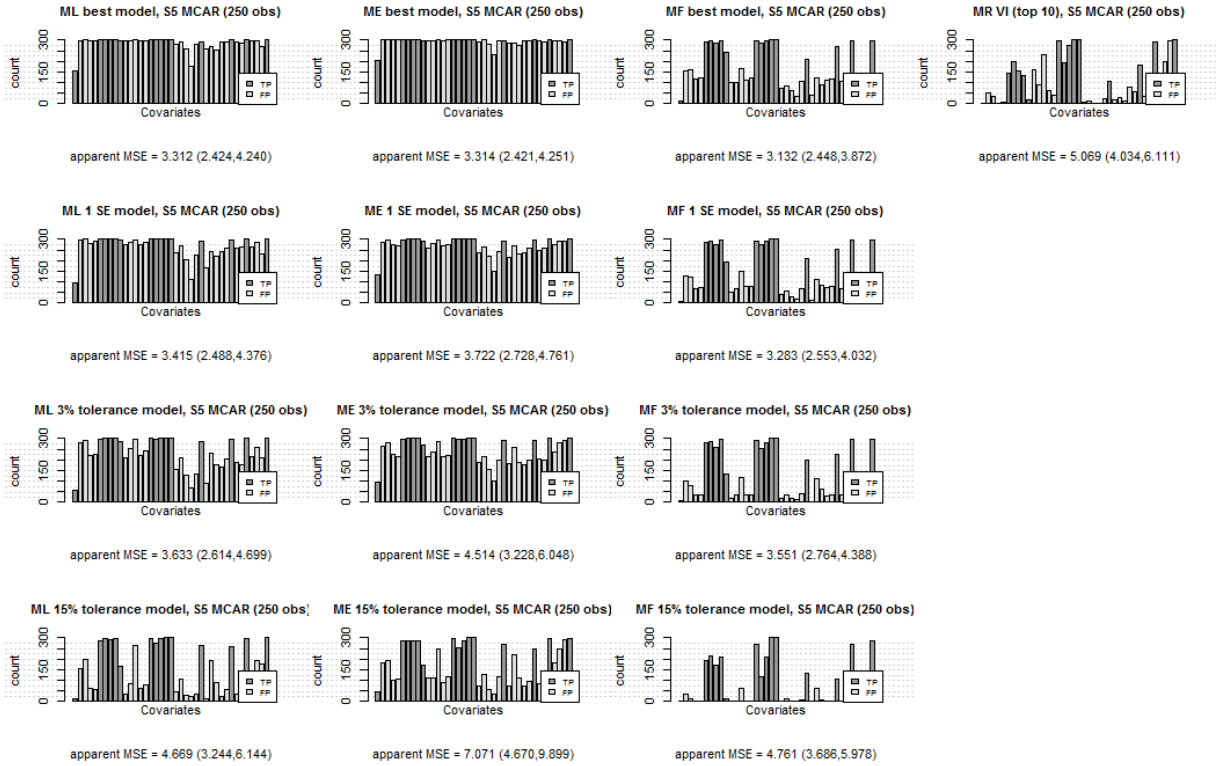


Figure 2.33: Comparison of **variable inclusion frequency** by 3 methods run on 300 simulated **20-covariate** datasets with **250 observations** for scenario **S4** with **MAR** data (assumption of moderation, complete outcome). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME), MissForest-Lasso (MF) and MissForest-Random Forests (MR). ML, ME and MF variable inclusion frequencies are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For MR only the best model is computed through bootstrap tuning of the parameter given by the number of variables chosen randomly at each split to build the trees and a variable is considered included in the model when its importance is among the top 10 variable importances.

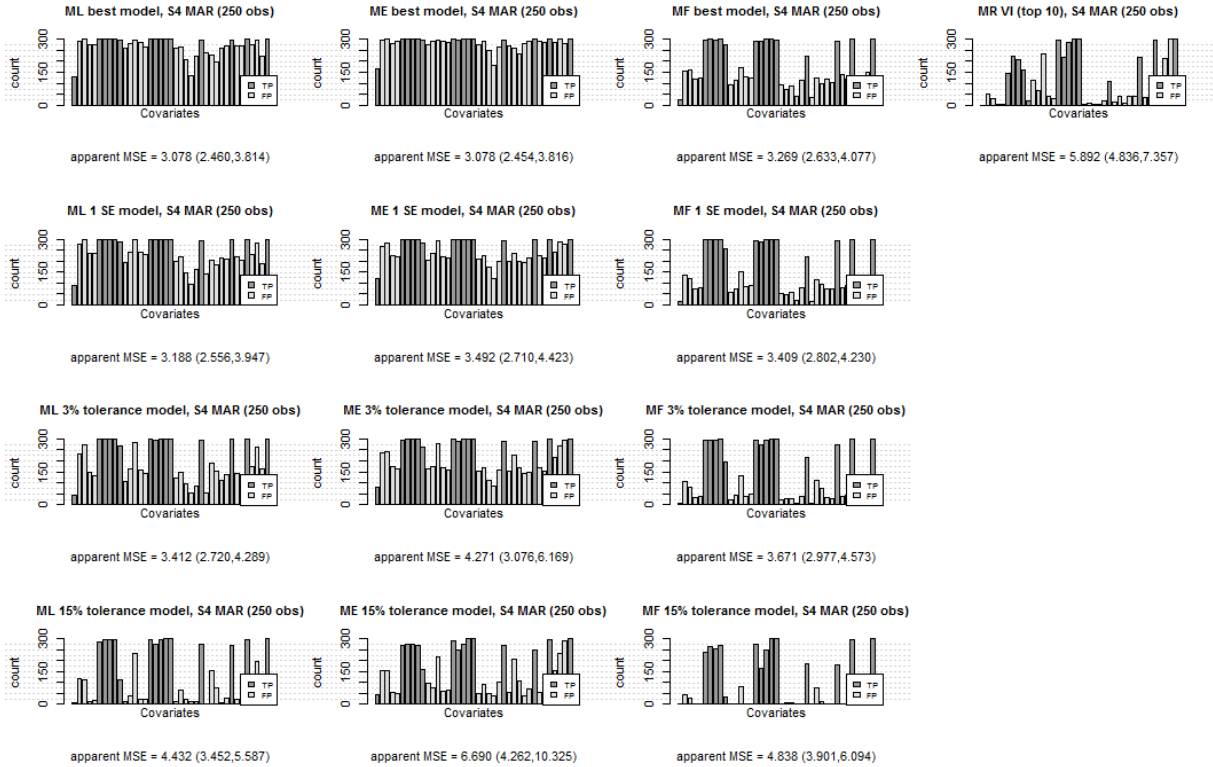
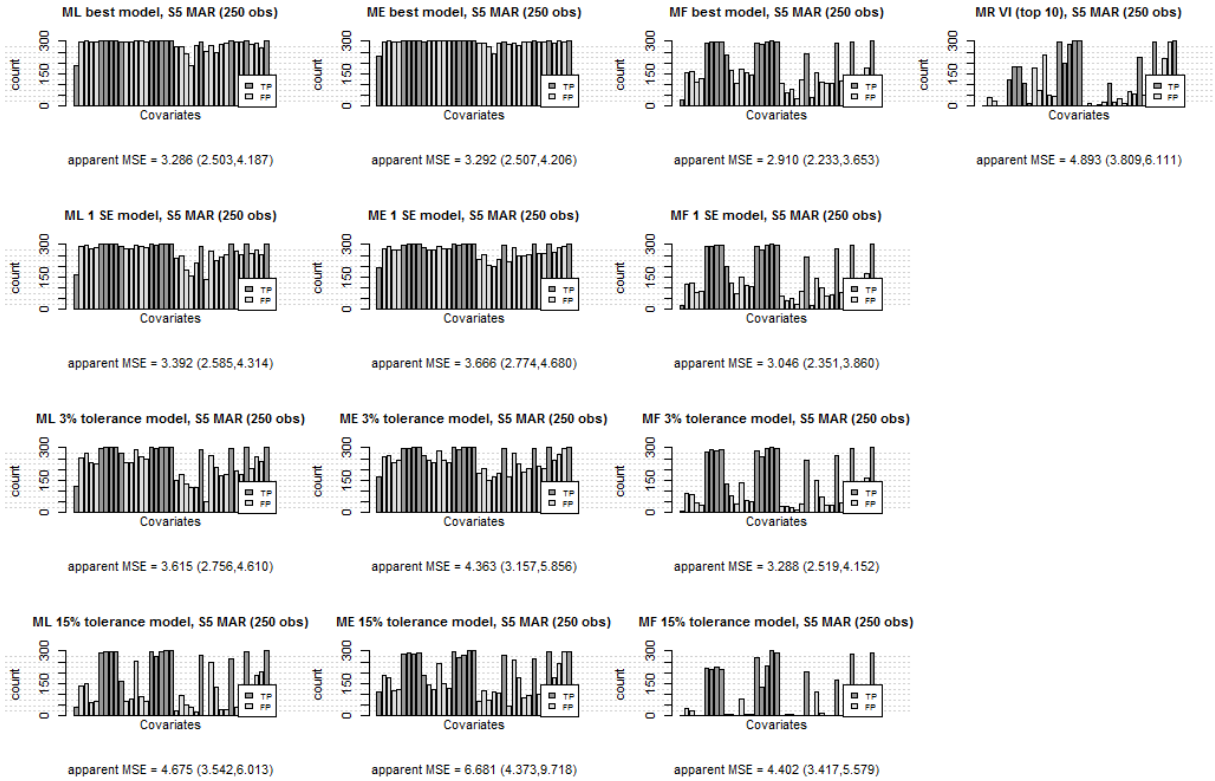


Figure 2.34: Comparison of **variable inclusion frequency** by 3 methods run on 300 simulated **20-covariate** datasets with **250 observations** for scenario **S5** with **MAR** data (assumption of moderation, missing data also in the outcome). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME), MissForest-Lasso (MF) and MissForest-Random Forests (MR). ML, ME and MF variable inclusion frequencies are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For MR only the best model is computed through bootstrap tuning of the parameter given by the number of variables chosen randomly at each split to build the trees and a variable is considered included in the model when its importance is among the top 10 variable importances.



### 2.3.2 Results from 100-covariate datasets, 15 true predictors

#### MICE-Lasso and MICE-Elasticnet: 100-covariate data results

**Lasso and Elasticnet S3: No missing data, Assumption of moderation** When the 100 covariates were weakly correlated ($\rho$**=0.2**) in absence of missing data and with moderators among the TPs (scenario S3), Lasso and Elasticnet best models had a similar good apparent performance. However, when correcting for optimism, the mean corrected MSE for the Lasso best model was 1.315 (2.5th and 97.5th percentiles being 1.101 and 1.493, see Table 2.29), within 30% of the theoretical MSE (1), and the corresponding corrected pseudo-$R^2$ was 0.742 (2.5th and 97.5th percentiles being 0.694 and 0.786, the mean variance of the simulated outcomes being 5.120 (SD 0.327)). With increasing penalty tolerance, the corrected MSEs were poor for the Lasso, even though the pseudo-$R^2$ for the 3% tolerance model was only 6% lower compared to the best model. The Elasticnet best model had a similar mean bootstrap-corrected MSE estimate (see Table 2.30), but the tolerance Elasticnet models accuracies got worse than Lasso given the same increasing $\lambda$ tolerance.

In this 100 covariate data scenario, the linear predictor had 234 FPs and only 15 TPs (see Subsection2.2.2). Despite the sparsity of the true vector of coefficients (6.4% non-zero entries), Lasso and Elasticnet best models had high sensitivity of selection by choosing approximately 89.0% (SD 5) of the TPs and 19.0% (SD 5) of the FPs. However, the large number of FPs in the model specification caused the best and the 1 SE tolerance models to always select more FPs than TPs (PPV of up to 32.2%, SD 7.1, see Table 2.31). With increasing penalty tolerance, the number of false positive rate decreased and only the 3% tolerance model of Lasso showed good variable selection performance. The exact true model and the true model but 1 TP were never selected. There was some general underfitting in the variable inclusion frequency for the best Lasso model, which increased with increasing tolerance (see Figure A.32).

When there was high correlation between covariates ($\rho$**=0.8**), prediction accuracy was slightly better compared to the weak correlation scenario (see tables A.10 and A.11 in the Appendix). In contrast, the variable selection was less precise (see Table 2.31). Lasso and Elasticnet tended to select more FPs as they were highly correlated with the TPs for an almost acceptable variable selection performance (3% tolerance model SEN being 70% (SD 14) and PPV being only 58% (SD 6)).

**MICE-Lasso and MICE-Elasticnet S5: Assumption of moderation, Missing data also in outcome (20% missingness MAR and MCAR** When there were missing data in both outcome and predictors and the predictors were weakly correlated ($\rho$**=0.2**), MICE-Lasso showed better prediction accuracy than MICE-Elasticnet (see tables 2.29 and 2.30). However, the average optimism-corrected MSE of the best MICE-Lasso model was very far from the theoretical MSE (2.879, 2.5th and 97.5th percentiles being 2.379 and 3.349, i.e 41% decrease in corrected pseudo-$R^2$ for MCAR data; 2.662, 2.5th and 97.5th percentiles being 2.220 and 3.227, for a 35% decrease in the pseudo-$R^2$ for MAR data) and it was much worsened compared to the complete data scenario, and the tolerance models estimates were also even poorer (see Figure 2.35). Average estimates of MSE internal and external optimism were far away from each other with the external estimate being smaller in absolute value (see Figures 2.39). When data were MAR, the bias due to resampling was reduced compared to MCAR because of the better accuracy.

On the other hand, the **0.8 correlation** setting accuracy results were closer to the complete data scenario S3 (see tables A.10 and A.11 in the Appendix) with the best MICE-Lasso and MICE-Elasticnet models showing a decrease in pseudo-$R^2$ of only approximately 6% for MCAR data and 10% for MAR data compared to scenario S3. Strongly correlated variables with MAR missing data were imputed in a less accurate way then MCAR missing data, contrarily to what happened in the low correlation case.

Alike in the 20-covariate case, MICE-Elasticnet had a better discriminative performance than Elasticnet in the higher tolerance models (see Figures 2.35 and 2.36). Moreover, mean optimism-corrected MSE estimates for both MICE-Lasso and MICE-Elasticnet seemed to be lower than Elasticnet estimates for the tolerance models, due to the decrease in mean internal optimism with increasing tolerance level.

The behaviour of MICE-Lasso and MICE-Elasticnet in variable selection (in both correlation scenarios) reflected the poor performance in the 20-covariate simulations (see section 2.3.1). Again all variables were selected most of the times without any preferences for TPs (see Table 2.31 and figures A.36, A.37, A.34 and A.35).

Table 2.29: **Accuracy** simulation study results for **MICE-Lasso** analysis with Harrell (1996) bootstrap validation: scenarios **S3** (assumption of moderation, complete data) and **S5** (assumption of moderation, missing data also in the outcome), based on 300 data sets of **100 variables** each (**n=500**) with between-covariate correlation of 0.2. Means of all estimates along with their corresponding 2.5th and 97.5th percentile values within parenthesis. The theoretical MSE is 1.

| **Estimates** | **Complete data** | | | |
|---|---|---|---|---|
| | Best model | 1 SE tolerance | 3% tolerance | 15% tolerance |
| $MSE_{apparent}$ | 0.951 (0.780,1.096) | 1.025 (0.841,1.173) | 1.366 (1.170,1.559) | 1.606 (1.380,1.823) |
| $\beta_{LP}$ | 1.093 (1.078,1.113) | 1.107 (1.088,1.130) | 1.190 (1.151,1.235) | 1.268 (1.203,1.355) |
| Tuning $\lambda$ | 0.048 (0.039,0.061) | 0.126 (0.107,0.149) | 0.060 (0.049,0.072) | 0.185 (0.149,0.234) |
| $MSE_{ext}$ | 1.125 (1.035,1.216) | 1.143 (1.053,1.238) | 1.349 (1.214,1.520) | 1.552 (1.365,1.746) |
| $Optimism_{ext}$ | -0.174 (-0.334,-0.027) | -0.118 (-0.285,0.038) | 0.017 (-0.149,0.194) | 0.054 (-0.134,0.246) |
| $Optimism_{int}$ | -0.365 (-0.413,-0.315) | -0.319 (-0.365,-0.273) | -0.187 (-0.228,-0.142) | -0.133 (-0.172,-0.095) |
| $MSE_{corrected}$ | 1.315 (1.101,1.493) | 1.343 (1.140,1.526) | 1.553 (1.352,1.765) | 1.739 (1.510,1.970) |
| $\beta_{LP^*}$ | 1.035 (1.026,1.045) | 1.053 (1.042,1.065) | 1.124 (1.100,1.149) | 1.175 (1.137,1.214) |
| **MCAR** | | | | |
| $MSE_{apparent}$ | 1.817 (1.503,2.143) | 1.812 (1.498,2.135) | 1.864 (1.543,2.178) | 2.051 (1.686,2.461) |
| $\beta_{LP}$ | 0.940 (0.904,0.973) | 0.949 (0.909,0.988) | 1.070 (1.007,1.139) | 1.171 (1.088,1.278) |
| Tuning $\lambda$ | 0.017 (0.016,0.019) | 0.050 (0.034,0.072) | 0.018 (0.016,0.024) | 0.091 (0.056,0.142) |
| $MSE_{ext}$ | 2.413 (2.055,2.821) | 2.374 (1.993,2.792) | 2.039 (1.802,2.342) | 1.998 (1.782,2.274) |
| $Optimism_{ext}$ | -0.596 (-1.110,-0.119) | -0.562 (-1.087,-0.094) | -0.176 (-0.646,0.254) | 0.053 (-0.385,0.473) |
| $Optimism_{int}$ | -1.062 (-1.332,-0.806) | -1.054 (-1.322,-0.802) | -0.859 (-1.086,-0.655) | -0.728 (-0.923,-0.547) |
| $MSE_{corrected}$ | 2.879 (2.379,3.349) | 2.866 (2.372,3.335) | 2.722 (2.264,3.185) | 2.779 (2.299,3.251) |
| $\beta_{LP^*}$ | 0.855 (0.817,0.901) | 0.858 (0.821,0.904) | 0.954 (0.922,0.987) | 1.032 (0.999,1.064) |
| **MAR** | | | | |
| $MSE_{apparent}$ | 1.661 (1.400,1.971) | 1.658 (1.400,1.964) | 1.727 (1.450,2.079) | 1.914 (1.569,2.369) |
| $\beta_{LP}$ | 0.950 (0.919,0.981) | 0.959 (0.922,0.994) | 1.071 (1.015,1.142) | 1.165 (1.081,1.273) |
| Tuning $\lambda$ | 0.016 (0.016,0.019) | 0.048 (0.032,0.070) | 0.018 (0.016,0.023) | 0.087 (0.054,0.131) |
| $MSE_{ext}$ | 2.468 (2.060,2.916) | 2.430 (2.020,2.868) | 2.082 (1.817,2.403) | 2.027 (1.809,2.332) |
| $Optimism_{ext}$ | -0.808 (-1.306,-0.396) | -0.772 (-1.270,-0.364) | -0.355 (-0.776,0.035) | -0.113 (-0.500,0.271) |
| $Optimism_{int}$ | -1.002 (-1.245,-0.789) | -0.996 (-1.237,-0.786) | -0.811 (-1.008,-0.644) | -0.685 (-0.852,-0.538) |
| $MSE_{corrected}$ | 2.662 (2.220,3.227) | 2.654 (2.216,3.213) | 2.537 (2.101,3.050) | 2.599 (2.129,3.156) |
| $\beta_{LP^*}$ | 0.859 (0.815,0.900) | 0.862 (0.818,0.901) | 0.953 (0.918,0.984) | 1.028 (0.990,1.059) |

Table 2.30: **Accuracy** simulation study results for **MICE-Elasticnet** analysis with Harrell (1996) bootstrap validation: scenarios **S3** (assumption of moderation, complete data) and **S5** (assumption of moderation, missing data also in the outcome), based on 300 data sets of **100 variables** each (**n=500**) with between-covariate correlation of 0.2. Means of all estimates along with their corresponding 2.5th and 97.5th percentile values within parenthesis. The theoretical MSE is $1$.

| **Estimates** | **Complete data** | | | |
|---|---|---|---|---|
| | Best model | 1 SE tolerance | 3% tolerance | 15% tolerance |
| $MSE_{apparent}$ | 0.951 (0.766,1.111) | 1.038 (0.847,1.209) | 1.981 (1.613,2.377) | 3.123 (2.414,3.914) |
| $\beta_{LP}$ | 1.095 (1.079,1.117) | 1.112 (1.092,1.138) | 1.447 (1.302,1.666) | 2.116 (1.668,2.853) |
| Tuning $\alpha$ | 0.900 (0.900,0.900) | 0.900 (0.900,0.900) | 0.900 (0.900,0.900) | 0.900 (0.900,0.900) |
| Tuning $\lambda$ | 0.053 (0.040,0.065) | 0.321 (0.220,0.455) | 0.068 (0.051,0.083) | 0.651 (0.455,0.941) |
| $MSE_{ext}$ | 1.127 (1.040,1.218) | 1.151 (1.064,1.242) | 1.912 (1.598,2.337) | 3.119 (2.347,3.978) |
| $Optimism_{ext}$ | -0.176 (-0.341,-0.012) | -0.113 (-0.272,0.042) | 0.069 (-0.161,0.318) | 0.004 (-0.403,0.375) |
| $Optimism_{int}$ | -0.361 (-0.412,-0.310) | -0.295 (-0.340,-0.247) | -0.088 (-0.121,-0.057) | -0.071 (-0.098,-0.042) |
| $MSE_{corrected}$ | 1.311 (1.084,1.503) | 1.332 (1.129,1.516) | 2.069 (1.701,2.455) | 3.194 (2.484,3.976) |
| $\beta_{LP^*}$ | 1.040 (1.031,1.049) | 1.068 (1.054,1.083) | 1.320 (1.225,1.448) | 1.638 (1.421,1.954) |
| | | **MCAR** | | |
| $MSE_{apparent}$ | 1.921 (1.563,2.262) | 1.891 (1.547,2.224) | 1.910 (1.549,2.296) | 2.119 (1.625,2.669) |
| $\beta_{LP}$ | 0.891 (0.838,0.938) | 0.919 (0.852,0.989) | 1.131 (0.996,1.322) | 1.308 (1.099,1.665) |
| Tuning $\alpha$ | 0.214 (0.125,0.345) | 0.764 (0.615,0.890) | 0.318 (0.170,0.500) | 0.869 (0.805,0.900) |
| Tuning $\lambda$ | 0.035 (0.032,0.047) | 0.220 (0.118,0.365) | 0.050 (0.034,0.079) | 0.426 (0.231,0.702) |
| $MSE_{ext}$ | 2.753 (2.215,3.384) | 2.617 (2.080,3.277) | 2.170 (1.901,2.519) | 2.213 (1.916,2.607) |
| $Optimism_{ext}$ | -0.832 (-1.507,-0.249) | -0.726 (-1.416,-0.150) | -0.260 (-0.849,0.249) | -0.094 (-0.586,0.396) |
| $Optimism_{int}$ | -1.258 (-1.546,-0.981) | -1.214 (-1.494,-0.954) | -0.934 (-1.151,-0.735) | -0.799 (-0.982,-0.636) |
| $MSE_{corrected}$ | 3.179 (2.379,3.349) | 3.105 (2.372,3.335) | 2.844 (2.264,3.185) | 2.918 (2.299,3.251) |
| $\beta_{LP^*}$ | 0.787 (0.745,0.835) | 0.804 (0.765,0.850) | 0.954 (0.918,0.990) | 1.058 (1.019,1.099) |
| | | **MAR** | | |
| $MSE_{apparent}$ | 1.740 (1.443,2.063) | 1.715 (1.434,2.024) | 1.770 (1.441,2.192) | 1.984 (1.512,2.611) |
| $\beta_{LP}$ | 0.905 (0.855,0.953) | 0.933 (0.871,1.005) | 1.138 (1.003,1.326) | 1.306 (1.092,1.633) |
| Tuning $\alpha$ | 0.219 (0.130,0.340) | 0.773 (0.605,0.890) | 0.318 (0.170,0.485) | 0.874 (0.785,0.900) |
| Tuning $\lambda$ | 0.034 (0.032,0.043) | 0.211 (0.109,0.344) | 0.048 (0.033,0.075) | 0.404 (0.203,0.675) |
| $MSE_{ext}$ | 2.823 (2.279,3.466) | 2.679 (2.137,3.339) | 2.201 (1.905,2.595) | 2.231 (1.930,2.682) |
| $Optimism_{ext}$ | -1.083 (-1.723,-0.574) | -0.964 (-1.653,-0.456) | -0.431 (-0.988,0.026) | -0.247 (-0.707,0.182) |
| $Optimism_{int}$ | -1.210 (-1.475,-0.975) | -1.175 (-1.426,-0.955) | -0.906 (-1.087,-0.746) | -0.776 (-0.930,-0.637) |
| $MSE_{corrected}$ | 2.950 (2.461,3.533) | 2.890 (2.421,3.441) | 2.675 (2.229,3.259) | 2.760 (2.215,3.509) |
| $\beta_{LP^*}$ | 0.787 (0.742,0.831) | 0.802 (0.757,0.841) | 0.942 (0.903,0.978) | 1.040 (0.991,1.083) |

Table 2.31: **Variable selection** simulation study results for scenarios **S3** (assumption of moderation, without missing data) and **S5** (assumption of moderation, missing data also in the outcome) for **MICE-Lasso** best and tolerance models in the case of **100 covariates** and 300 samples of 500 observations, between-covariate correlation of 0.2 and 0.8. The following results are shown: the estimated percentages of the times all the 10 true predictors (TP) are selected at the same time, the estimated percentages of the times the true model is selected apart from one variable, the percentages of the times the TPs are the top ranked variables among the selected, the percentage of times the TP are the top ranked variables apart from one TP, the average percentages of selected TP among the TP (sensitivity, SEN), the average percentages of selected false positive predictors (FP) among FP (false positive rate, FPR), and the average percentages of selected TP among the selected variables (positive predictive value, PPV). The mean percentages are shown along with their standard deviation (SD)

| | **LASSO** | | | | **MICE-LASSO** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Variable selection** | **Complete data** | | | | **MCAR** | | | | **MAR** | | | |
| | Best | 1 SE tolerance | 3% tolerance | 15% tolerance | Best | 1 SE tolerance | 3% tolerance | 15% tolerance | Best | 1 SE tolerance | 3% tolerance | 15% tolerance |
| | | | | | | **correlation = 0.2** | | | | | | |
| **% true models** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **% true models but 1** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **% TP in top 10** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **% TP in top 10 but 1** | 9.3 | 5.3 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **SEN (SD)** | 88.9 (4.8) | 86.0 (5.8) | 74.8 (6.2) | 63.2 (8.2) | 94.7 (3.0) | 94.6 (3.0) | 90.0 (4.7) | 83.7 (6.5) | 94.0 (2.4) | 93.8 (2.6) | 89.7 (4.7) | 83.6 (7.0) |
| **FPR (SD)** | 18.6 (4.5) | 13.2 (3.7) | 2.1 (1.3) | 0.8 (0.5) | 100.0 (0.3) | 100.0 (0.4) | 94.7 (2.6) | 83.2 (6.8) | 100.0 (0.4) | 100.0 (0.4) | 94.3 (2.7) | 82.7 (6.7) |
| **PPV (SD)** | 24.7 (4.5) | 31.3 (6.1) | 76.4 (13.0) | 92.8 (8.8) | 5.7 (0.2) | 5.7 (0.2) | 5.8 (0.3) | 6.1 (0.6) | 5.7 (0.1) | 5.7 (0.1) | 5.8 (0.3) | 6.1 (0.6) |
| | | | | | | **correlation = 0.8** | | | | | | |
| **% true models** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **% true models but 1** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **% TP in top 10** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **% TP in top 10 but 1** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **SEN (SD)** | 71.9 (6.2) | 68.7 (6.0) | 57.9 (5.7) | 51.0 (6.7) | 93.5 (4.9) | 92.7 (5.1) | 81.5 (5.2) | 77.1 (5.2) | 93.5 (4.3) | 92.5 (4.4) | 82.0 (5.4) | 76.8 (6.3) |
| **FPR (SD)** | 11.6 (3.6) | 8.3 (2.8) | 2.2 (1.2) | 1.2 (0.7) | 97.5 (1.1) | 96.7 (1.3) | 74.1 (6.9) | 47.1 (9.1) | 97.3 (1.3) | 96.3 (1.6) | 68.1 (8.5) | 39.3 (9.2) |
| **PPV (SD)** | 30.5 (6.2) | 37.5 (8.0) | 70.1 (13.6) | 82.9 (13.4) | 5.8 (0.3) | 5.8 (0.3) | 6.7 (0.7) | 9.9 (1.8) | 5.8 (0.2) | 5.8 (0.3) | 7.3 (0.9) | 11.7 (2.4) |
| | **ELASTICNET** | | | | **MICE-ELASTICNET** | | | | | | | |
| **Variable selection** | **Complete data** | | | | **MCAR** | | | | **MAR** | | | |
| | Best | 1 SE tolerance | 3% tolerance | 15% tolerance | Best | 1 SE tolerance | 3% tolerance | 15% tolerance | Best | 1 SE tolerance | 3% tolerance | 15% tolerance |
| | | | | | | **correlation = 0.2** | | | | | | |
| **% true models** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **% true models but 1** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **% TP in top 10** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **% TP in top 10 but 1** | 10.0 | 3.3 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **SEN (SD)** | 89.0 (4.8) | 85.8 (5.8) | 53.1 (4.9) | 36.0 (8.0) | 99.2 (2.2) | 99.0 (2.4) | 95.3 (3.9) | 92.8 (4.7) | 99.0 (2.3) | 98.7 (2.7) | 94.7 (3.5) | 92.1 (4.8) |
| **FPR (SD)** | 18.8 (4.6) | 12.8 (3.9) | 0.5 (0.1) | 0.4 (0.0) | 100.0 (0.0) | 100.0 (0.0) | 99.7 (1.7) | 96.6 (5.4) | 100.0 (0.0) | 100.0 (0.0) | 99.5 (1.7) | 95.9 (5.6) |
| **PPV (SD)** | 24.5 (4.7) | 32.2 (7.1) | 99.0 (3.2) | 100.0 (0.7) | 6.0 (0.1) | 6.0 (0.1) | 5.8 (0.2) | 5.8 (0.4) | 6.0 (0.1) | 5.9 (0.2) | 5.8 (0.2) | 5.8 (0.4) |
| | | | | | | **correlation = 0.8** | | | | | | |
| **% true models** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **% true models but 1** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **% TP in top 10** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **% TP in top 10 but 1** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **SEN (SD)** | 72.3 (5.9) | 67.2 (6.3) | 40.3 (5.8) | 31.3 (6.4) | 99.6 (1.5) | 99.1 (2.4) | 92.0 (5.2) | 87.2 (5.5) | 99.2 (2.2) | 98.5 (2.9) | 91.4 (5.3) | 86.9 (5.8) |
| **FPR (SD)** | 12.1 (3.8) | 7.0 (2.4) | 0.6 (0.3) | 0.5 (0.2) | 100.0 (0.1) | 100.0 (0.3) | 93.7 (5.5) | 79.9 (10.9) | 100.0 (0.3) | 100.0 (0.9) | 90.5 (7.5) | 73.2 (12.6) |
| **PPV (SD)** | 29.7 (6.4) | 41.3 (8.4) | 93.9 (9.7) | 97.9 (6.0) | 6.0 (0.1) | 6.0 (0.1) | 6.0 (0.4) | 6.7 (0.9) | 6.0 (0.1) | 6.0 (0.2) | 6.1 (0.5) | 7.3 (1.3) |

**MissForest-Lasso and MissForest-Elasticnet: 100-covariate data results**

When missing data in predictors and outcome were present, MissForest outperformed MICE as an imputation method when combined with Lasso in both low and high correlation scenarios. MissForest-Lasso discrimination estimates were still poor compared to the theoretical MSE, but closer to the complete data scenario S3 accuracy performance than MICE-Lasso (see tables 2.32 and A.15 and figures 2.35 and 2.36). Also estimates of internal and external optimism were smaller in absolute value for MissForest-Lasso than MICE-Lasso (see Figures 2.39 and 2.40) and the differences between MAR and MCAR data results were now reduced. Instead, MissForest-Elasticnet tolerance models enormously underfitted the data such that the 300 runs estimates of accuracy were mostly missing and unstable (see Table 2.33 and A.16). Only MissForest-Lasso calibration performance for the low correlation scenario was poorer with respect to MICE-Lasso (see Figures 2.37 and 2.38).

Variable selection for MissForest-Lasso was generally poor in the low correlation scenario, but much better than MICE-Lasso performance (see Table 2.34 and compare figures A.32, A.33 with A.36,A.37, A.34, A.35). As in the other scenarios, the sensitivity of selection (SEN) decreased with increasing penalty strength and the opposite happened to the positive predictive value (PPV). All the MissForest-Lasso models did not have high SEN in both correlation scenarios (up to 66.1%, SD 6.7, in the best models and up to 46.2%, SD 7.9, in the 3% tolerance models) and only the 3% and 15% tolerance models in the low correlation scenario had high PPV (83.5%, SD 14.2 and 94.7%, SD 8.9) respectively, see Table 2.34).

The pattern of results between weak and strong correlation settings repeated: accuracy estimates for the 0.8 correlation scenario were poor according to my subjective criteria, but more precise and closer to the complete data scenario estimates (within 50% to 70% of the theoretical MSE) than the 0.2 correlation scenario; while on average the tolerance models had a lower PPV (56% SD 13 in the 3% tolerance MissForest-Lasso model) in the same high correlation scenario compared to the weak correlation case.

Table 2.32: **Accuracy** simulation study results for **MissForest-Lasso** analysis with Harrell (1996) bootstrap validation: scenarios **S3** (assumption of moderation, complete data) and **S5** (assumption of moderation, missing data also in the outcome), based on 300 data sets of **100 variables** each (**n=500**) with between-covariate correlation of 0.2. Means of all estimates along with their corresponding 2.5th and 97.5th percentile values within parenthesis. The theoretical MSE is $1$.

| **Estimates** | **Complete data** | | | |
|---|---|---|---|---|
| | Best model | 1 SE tolerance | 3% tolerance | 15% tolerance |
| $\text{MSE}_{apparent}$ | 0.951 (0.780,1.096) | 1.025 (0.841,1.173) | 1.366 (1.170,1.559) | 1.606 (1.380,1.823) |
| $\beta_{LP}$ | 1.093 (1.078,1.113) | 1.107 (1.088,1.130) | 1.190 (1.151,1.235) | 1.268 (1.203,1.355) |
| Tuning $\lambda$ | 0.048 (0.039,0.061) | 0.126 (0.107,0.149) | 0.060 (0.049,0.072) | 0.185 (0.149,0.234) |
| $\text{MSE}_{ext}$ | 1.125 (1.035,1.216) | 1.143 (1.053,1.238) | 1.349 (1.214,1.520) | 1.552 (1.365,1.746) |
| $\text{Optimism}_{ext}$ | -0.174 (-0.334,-0.027) | -0.118 (-0.285,0.038) | 0.017 (-0.149,0.194) | 0.054 (-0.134,0.246) |
| $\text{Optimism}_{int}$ | -0.365 (-0.413,-0.315) | -0.319 (-0.365,-0.273) | -0.187 (-0.228,-0.142) | -0.133 (-0.172,-0.095) |
| $\text{MSE}_{corrected}$ | 1.315 (1.101,1.493) | 1.343 (1.140,1.526) | 1.553 (1.352,1.765) | 1.739 (1.510,1.970) |
| $\beta_{LP^*}$ | 1.035 (1.026,1.045) | 1.053 (1.042,1.065) | 1.124 (1.100,1.149) | 1.175 (1.137,1.214) |
| | | **MCAR** | | |
| $\text{MSE}_{apparent}$ | 1.739 (1.409,2.071) | 1.860 (1.547,2.192) | 2.333 (1.973,2.689) | 2.674 (2.259,3.079) |
| $\beta_{LP}$ | 1.168 (1.131,1.215) | 1.203 (1.155,1.265) | 1.470 (1.335,1.679) | 1.764 (1.529,2.173) |
| Tuning $\lambda$ | 0.087 (0.068,0.459) | 0.114 (0.068,0.459) | 0.275 (0.068,0.459) | 0.405 (0.068,0.459) |
| $\text{MSE}_{ext}$ | 1.713 (1.500,1.981) | 1.776 (1.546,2.082) | 2.354 (1.972,2.778) | 2.872 (2.359,3.431) |
| $\text{Optimism}_{ext}$ | 0.025 (-0.339,0.400) | 0.084 (-0.288,0.463) | -0.021 (-0.470,0.421) | -0.198 (-0.739,0.320) |
| $\text{Optimism}_{int}$ | -0.625 (-0.762,-0.497) | -0.516 (-0.644,-0.397) | -0.218 (-0.321,-0.130) | -0.128 (-0.221,-0.057) |
| $\text{MSE}_{corrected}$ | 2.363 (1.947,2.781) | 2.376 (1.977,2.786) | 2.551 (2.159,2.936) | 2.802 (2.361,3.231) |
| $\beta_{LP^*}$ | 1.026 (1.003,1.052) | 1.068 (1.045,1.097) | 1.252 (1.188,1.336) | 1.443 (1.322,1.640) |
| | | **MAR** | | |
| $\text{MSE}_{apparent}$ | 1.699 (1.441,2.000) | 1.816 (1.559,2.128) | 2.281 (1.955,2.632) | 2.619 (2.241,3.016) |
| $\beta_{LP}$ | 1.165 (1.127,1.213) | 1.197 (1.148,1.256) | 1.460 (1.318,1.683) | 1.772 (1.524,2.224) |
| Tuning $\lambda$ | 0.086 (0.068,0.459) | 0.112 (0.068,0.459) | 0.280 (0.068,0.459) | 0.423 (0.068,0.459) |
| $\text{MSE}_{ext}$ | 1.831 (1.581,2.180) | 1.900 (1.631,2.215) | 2.513 (2.112,3.017) | 3.025 (2.547,3.610) |
| $\text{Optimism}_{ext}$ | -0.131 (-0.535,0.258) | -0.084 (-0.480,0.271) | -0.233 (-0.756,0.187) | -0.405 (-0.998,0.025) |
| $\text{Optimism}_{int}$ | -0.608 (-0.725,-0.497) | -0.503 (-0.606,-0.405) | -0.213 (-0.301,-0.126) | -0.124 (-0.207,-0.052) |
| $\text{MSE}_{corrected}$ | 2.307 (1.967,2.670) | 2.318 (1.991,2.696) | 2.494 (2.151,2.852) | 2.743 (2.378,3.139) |
| $\beta_{LP^*}$ | 1.026 (1.000,1.052) | 1.066 (1.037,1.098) | 1.243 (1.173,1.327) | 1.438 (1.304,1.612) |

Table 2.33: **Accuracy** simulation study results for **MissForest-Elasticnet** analysis with Harrell (1996) bootstrap validation: scenarios **S3** (assumption of moderation, complete data) and **S5** (assumption of moderation, missing data also in the outcome), based on 300 data sets of **100 variables** each (**n=500**) with between-covariate correlation of 0.2. Means of all estimates along with their corresponding 2.5th and 97.5th percentile values within parenthesis. The theoretical MSE is $1$.

| **Estimates** | **Complete data** | | | |
|---|---|---|---|---|
| | Best model | 1 SE tolerance | 3% tolerance | 15% tolerance |
| $\text{MSE}_{apparent}$ | 0.951 (0.766,1.111) | 1.038 (0.847,1.209) | 1.981 (1.613,2.377) | 3.123 (2.414,3.914) |
| $\beta_{LP}$ | 1.095 (1.079,1.117) | 1.112 (1.092,1.138) | 1.447 (1.302,1.666) | 2.116 (1.668,2.853) |
| Tuning $\alpha$ | 0.900 (0.900,0.900) | 0.900 (0.900,0.900) | 0.900 (0.900,0.900) | 0.900 (0.900,0.900) |
| Tuning $\lambda$ | 0.053 (0.040,0.065) | 0.321 (0.220,0.455) | 0.068 (0.051,0.083) | 0.651 (0.455,0.941) |
| $\text{MSE}_{ext}$ | 1.127 (1.040,1.218) | 1.151 (1.064,1.242) | 1.912 (1.598,2.337) | 3.119 (2.347,3.978) |
| $\text{Optimism}_{ext}$ | -0.176 (-0.341,-0.012) | -0.113 (-0.272,0.042) | 0.069 (-0.161,0.318) | 0.004 (-0.403,0.375) |
| $\text{Optimism}_{int}$ | -0.361 (-0.412,-0.310) | -0.295 (-0.340,-0.247) | -0.088 (-0.121,-0.057) | -0.071 (-0.098,-0.042) |
| $\text{MSE}_{corrected}$ | 1.311 (1.084,1.503) | 1.332 (1.129,1.516) | 2.069 (1.701,2.455) | 3.194 (2.484,3.976) |
| $\beta_{LP^*}$ | 1.040 (1.031,1.049) | 1.068 (1.054,1.083) | 1.320 (1.225,1.448) | 1.638 (1.421,1.954) |
| | | **MCAR** | | |
| $\text{MSE}_{apparent}$ | 1.740 (1.424,2.059) | 2.132 (1.752,2.623) | 4.123 (3.482,4.832) | 4.258 (3.678,4.954) |
| $\beta_{LP}$ | 1.172 (1.132,1.224) | 1.344 (1.211,1.547) | 20.555 (3.134,133.146) | 203.357 (6.282,566.300) |
| Tuning $\lambda$ | 0.101 (0.065,0.135) | 0.232 (0.135,0.357) | 1.389 (0.941,1.947) | 2.289 (1.528,3.162) |
| $\text{MSE}_{ext}$ | 1.715 (1.498,1.985) | 2.072 (1.661,2.615) | 5.010 (4.077,5.258) | 5.207 (5.196,5.265) |
| $\text{Optimism}_{ext}$ | 0.024 (-0.372,0.378) | 0.061 (-0.361,0.445) | -0.887 (-1.529,-0.187) | -0.949 (-1.535,-0.254) |
| $\text{Optimism}_{int}$ | -0.664 (-0.812,-0.526) | -0.418 (-0.608,-0.273) | -0.178 (-0.320,-0.068) | -0.116 (-0.220,-0.030) |
| $\text{MSE}_{corrected}$ | 2.403 (1.968,2.825) | 2.550 (2.022,3.159) | 4.301 (3.651,4.986) | 4.374 (3.764,5.101) |
| $\beta_{LP^*}$ | 1.034 (1.013,1.062) | 1.176 (1.112,1.248) | 2.811 (1.530,7.513) | 9.066 (2.107,26.418) |
| | | **MAR** | | |
| $\text{MSE}_{apparent}$ | 1.696 (1.410,2.007) | 2.093 (1.722,2.462) | 3.968 (3.280,4.614) | 4.111 (3.388,4.639) |
| $\beta_{LP}$ | 1.168 (1.131,1.220) | 1.344 (1.213,1.588) | NA (2.836,93.107) | NA (8.821,29.264) |
| Tuning $\lambda$ | 0.100 (0.065,0.135) | 0.241 (0.135,0.357) | 1.417 (1.199,1.947) | 2.373 (1.947,3.162) |
| $\text{MSE}_{ext}$ | 1.833 (1.582,2.137) | 2.236 (1.809,2.760) | 4.997 (4.054,5.241) | 5.204 (5.196,5.264) |
| $\text{Optimism}_{ext}$ | -0.137 (-0.548,0.229) | -0.143 (-0.670,0.242) | -1.029 (-1.727,-0.426) | -1.093 (-1.727,-0.553) |
| $\text{Optimism}_{int}$ | -0.650 (-0.778,-0.529) | -0.429 (-0.592,-0.282) | -0.193 (-0.315,-0.075) | -0.126 (-0.224,-0.037) |
| $\text{MSE}_{corrected}$ | 2.346 (2.002,2.742) | 2.522 (2.074,3.029) | 4.160 (3.471,4.795) | 4.237 (3.536,4.798) |
| $\beta_{LP^*}$ | 1.035 (1.011,1.064) | 1.162 (1.096,1.236) | 2.634 (1.423,7.384) | 50.914 (1.862,19.458) |

Table 2.34: **Variable selection** simulation study results for scenarios **S3** (assumption of moderation, without missing data) and **S5** (assumption of moderation, missing data also in the outcome) for **MissForest-Lasso** best and tolerance models in the case of **100 covariates** and 300 samples of 500 observations, between-covariate correlation of 0.2 and 0.8. The following results are shown: the estimated percentages of the times all the 10 true predictors (TP) are selected at the same time, the estimated percentages of the times the true model is selected apart from one variable, the percentages of the times the TPs are the top ranked variables among the selected, the percentage of times the TP are the top ranked variables apart from one TP, the average percentages of selected TP among the TP (sensitivity, SEN), the average percentages of selected false positive predictors (FP) among FP (false positive rate, FPR), and the average percentages of selected TP among the selected variables (positive predictive value, PPV). The mean percentages are shown along with their standard deviation (SD). The asterisk is present when the means and standard deviations were computed by removing the missing performances.

| | **LASSO** | | | | **MissForest-LASSO** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Variable selection** | **Complete data** | | | | **MCAR** | | | | **MAR** | | | |
| | Best | 1 SE tolerance | 3% tolerance | 15% tolerance | Best | 1 SE tolerance | 3% tolerance | 15% tolerance | Best | 1 SE tolerance | 3% tolerance | 15% tolerance |
| | | | | | | **correlation = 0.2** | | | | | | |
| **% true models** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **% true models but 1** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **% TP in top 10** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **% TP in top 10 but 1** | 9.3 | 5.3 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **SEN (SD)** | 88.9 (4.8) | 86.0 (5.8) | 74.8 (6.2) | 63.2 (8.2) | 66.1 (6.7) | 61.6 (6.4) | 45.4 (6.2) | 36.7 (6.5) | 64.0 (6.4) | 60.5 (6.3) | 42.2 (8.0) | 32.0 (8.0) |
| **FPR (SD)** | 18.6 (4.5) | 13.2 (3.7) | 2.1 (1.3) | 0.8 (0.5) | 11.5 (4.2) | 7.0 (2.9) | 1.1 (0.8) | 0.6 (0.3) | 11.7 (4.1) | 7.2 (3.0) | 1.2 (0.8) | 0.6 (0.3) |
| **PPV (SD)** | 24.7 (4.5) | 31.3 (6.1) | 76.4 (13.0) | 92.8 (8.8) | 29.6 (7.3) | 40.4 (10.4) | 83.5 (14.2) | 94.7 (8.9) | 28.3 (7.0) | 39.0 (10.1) | 81.2 (14.9) | 93.7 (10.5) |
| | | | | | | **correlation = 0.8** | | | | | | |
| **% true models** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **% true models but 1** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **% TP in top 10** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **% TP in top 10 but 1** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **SEN (SD)** | 71.9 (6.2) | 68.7 (6.0) | 57.9 (5.7) | 51.0 (6.7) | 64.9 (6.6) | 62.3 (7.0) | 46.2 (7.9) | 38.2 (7.5) | 64.8 (6.6) | 62.3 (7.1) | 48.2 (7.2) | 37.4 (7.6) |
| **FPR (SD)** | 11.6 (3.6) | 8.3 (2.8) | 2.2 (1.2) | 1.2 (0.7) | 11.6 (3.3) | 8.5 (2.5) | 3.0 (1.3) | 1.8 (1.0) | 11.6 (3.0) | 8.7 (2.3) | 3.1 (1.4) | 1.8 (0.9) |
| **PPV (SD)** | 30.5 (6.2) | 37.5 (8.0) | 70.1 (13.6) | 82.9 (13.4) | 28.3 (6.2) | 34.3 (7.0) | 56.3 (13.0) | 67.1 (17.1) | 28.1 (5.6) | 33.7 (6.8) | 56.1 (13.6) | 66.9 (16.0) |
| | **ELASTICNET** | | | | **MissForest-ELASTICNET** | | | | | | | |
| **Variable selection** | **Complete data** | | | | **MCAR** | | | | **MAR** | | | |
| | Best | 1 SE tolerance | 3% tolerance | 15% tolerance | Best | 1 SE tolerance | 3% tolerance | 15% tolerance | Best | 1 SE tolerance | 3% tolerance | 15% tolerance |
| | | | | | | **correlation = 0.2** | | | | | | |
| **% true models** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **% true models but 1** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **% TP in top 10** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **% TP in top 10 but 1** | 10.0 | 3.3 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **SEN (SD)** | 89.0 (4.8) | 85.8 (5.8) | 53.1 (4.9) | 36.0 (8.0) | 66.7 (6.5) | 53.1 (6.4) | 3.2 (4.7) | 0.1 (0.9) | 64.0 (6.4) | 60.5 (6.3) | 42.2 (8.0) | 32.0 (8.0) |
| **FPR (SD)** | 18.8 (4.6) | 12.8 (3.9) | 0.5 (0.1) | 0.4 (0.0) | 11.8 (4.1) | 2.6 (1.6) | 0.4 (0.0) | 0.4 (0.0) | 11.7 (4.1) | 7.2 (3.0) | 1.2 (0.8) | 0.6 (0.3) |
| **PPV (SD)** | 24.5 (4.7) | 32.2 (7.1) | 99.0 (3.2) | 100.0 (0.7) | 29.0 (7.3) | 66.3 (16.3) | 100.0* (0.0) | 100.0* (0.0) | 28.3 (7.0) | 39.0 (10.1) | 81.2 (14.9) | 93.7 (10.5) |
| | | | | | | **correlation = 0.8** | | | | | | |
| **% true models** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **% true models but 1** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **% TP in top 10** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **% TP in top 10 but 1** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **SEN (SD)** | 72.3 (5.9) | 67.2 (6.3) | 40.3 (5.8) | 31.3 (6.4) | 65.8 (7.0) | 56.3 (8.6) | 31.1 (8.1) | 13.0 (11.8) | 65.3 (7.0) | 57.6 (7.3) | 29.0 (6.9) | 11.2 (10.2) |
| **FPR (SD)** | 12.1 (3.8) | 7.0 (2.4) | 0.6 (0.3) | 0.5 (0.2) | 12.3 (3.8) | 5.5 (2.3) | 2.2 (2.2) | 1.4 (2.6) | 12.1 (3.4) | 5.8 (1.7) | 1.7 (1.0) | 0.8 (0.9) |
| **PPV (SD)** | 29.7 (6.4) | 41.3 (8.4) | 93.9 (9.7) | 97.9 (6.0) | 27.3 (5.8) | 43.5 (9.0) | 62.8 (22.3) | 72.8* (29.0) | 27.4 (5.4) | 41.9 (8.4) | 64.9 (18.9) | 79.0* (24.4) |

**MissForest-Conditional Random Forests: 100-covariate data results**

**Conditional Random Forests S3: No missing data, Assumption of moderation** Conditional RF had the worst prediction accuracy among the methods best models when data were complete: optimism-corrected MSE=1.897 (2.5th and 97.5th percentiles being 1.716 and 2.095, pseudo $R^2$=0.629, 0.581-0.667) for the low correlation scenario and 1.376 (1.215-1.553, pseudo $R^2$=0.731, 0.681-0.770) for the high correlation scenario (see Table 2.35). Also both MSE optimism internal and external were larger in absolute value, but close to each other (see Figure 2.39 and 2.40). Conditional RF accuracy was better with strongly correlated features than weakly correlated features like the other methods.

Despite the worst prediction accuracy result, the Conditional RF model always had the 15 TPs as top 15 features in the variable importance rank (see Figures A.32 and A.33), showing the best variable selection performance among the analysed methods.

**MissForest-Conditional RF S5: Assumption of moderation, Missing data also in outcome (20% missingness MAR and MCAR** MissForest-Conditional RF had poor discrimination similar to MissForest-Lasso, but had the best variable selection performance for scenario S5 (see Figures 2.35, 2.36, A.36, A.34, A.33, A.35, and Table 2.35). This missing data case accuracy result was better than the Conditional RF complete data scenario result. However, the mean calibration slope was the highest among the other methods best models calibration slopes in the low correlation scenario. Moreover, the estimated mean MSE internal and external optimisms were the largest in absolute value among the methods and were not close to each other in both correlation cases.

Table 2.35: **Accuracy** simulation study results for **MissForest-Conditional RF** analysis with Harrell bootstrap validation: scenarios **S3** (assumption of moderation, complete data) and **S5** (assumption of moderation, missing data also in the outcome), based on 300 data sets of **100 variables** each (n=500). Means of all estimates along with their corresponding 2.5th and 97.5th percentile values within parenthesis. The theoretical MSE is 1.

| **Estimates** | **correlation = 0.2** | | | **correlation = 0.8** | | |
|---|---|---|---|---|---|---|
| | **Complete data** | **MCAR** | **MAR** | **Complete data** | **MCAR** | **MAR** |
| $\text{MSE}_{apparent}$ | 1.256 (1.133,1.405) | 1.351 (1.167,1.561) | 1.290 (1.098,1.502) | 0.996 (0.870,1.141) | 0.956 (0.825,1.106) | 0.974 (0.844,1.132) |
| $\beta_{LP}$ | 1.257 (1.133,1.405) | 1.318 (1.167,1.561) | 1.298 (1.098,1.502) | 1.107 (0.870,1.141) | 1.107 (0.825,1.106) | 1.105 (0.844,1.132) |
| $\text{MSE}_{ext}$ | 2.102 (1.932,2.324) | 2.771 (2.458,3.132) | 2.102 (6.271,8.086) | 1.382 (1.271,1.506) | 1.656 (1.472,1.898) | 1.655 (1.488,1.866) |
| $\text{Optimism}_{ext}$ | -0.847 (-1.093,-0.605) | -1.419 (-1.841,-1.027) | -1.605 (-2.073,-1.179) | -0.386 (-0.565,-0.227) | -0.701 (2.774,-0.472) | -0.681 (-0.912,-0.449) |
| $\text{Optimism}_{int}$ | -0.642 (-0.707,-0.570) | -0.931 (-1.084,-0.784) | -0.889 (-1.032,-0.744) | -0.380 (-0.429,-0.332) | -0.487 (-0.564,-0.415) | -0.499 (-0.582,-0.421) |
| $\text{MSE}_{corrected}$ | 1.897 (1.716,2.095) | 2.283 (1.971,2.623) | 2.179 (1.874,2.507) | 1.376 (1.215,1.553) | 1.443 (1.255,1.653) | 1.472 (1.287,1.670) |
| $\beta_{LP^*}$ | 1.218 (1.170,1.268) | 1.239 (1.172,1.333) | 1.218 (1.162,1.291) | 1.094 (1.072,1.118) | 1.080 (1.062,1.105) | 1.081 (1.059,1.107) |

Figure 2.35: **Optimism-corrected MSE** estimates from 5 methods run on 300 simulated **100-covariate** datasets (**correlation 0.2**) with **500 observations** for scenarios S3 (assumption of moderation, without missing data) and S5 (assumption of moderation, missing data also in the outcome). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME), MissForest-Lasso (MFL), MissForest-Elasticnet (MFE), MissForest-Conditional RF (MC). ML, ME, MFL and MFE estimated MSEs are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S3 (first plot from the left), the Lasso (L), Elasticnet (E) and Conditional RF (CF) corrected MSEs are shown.

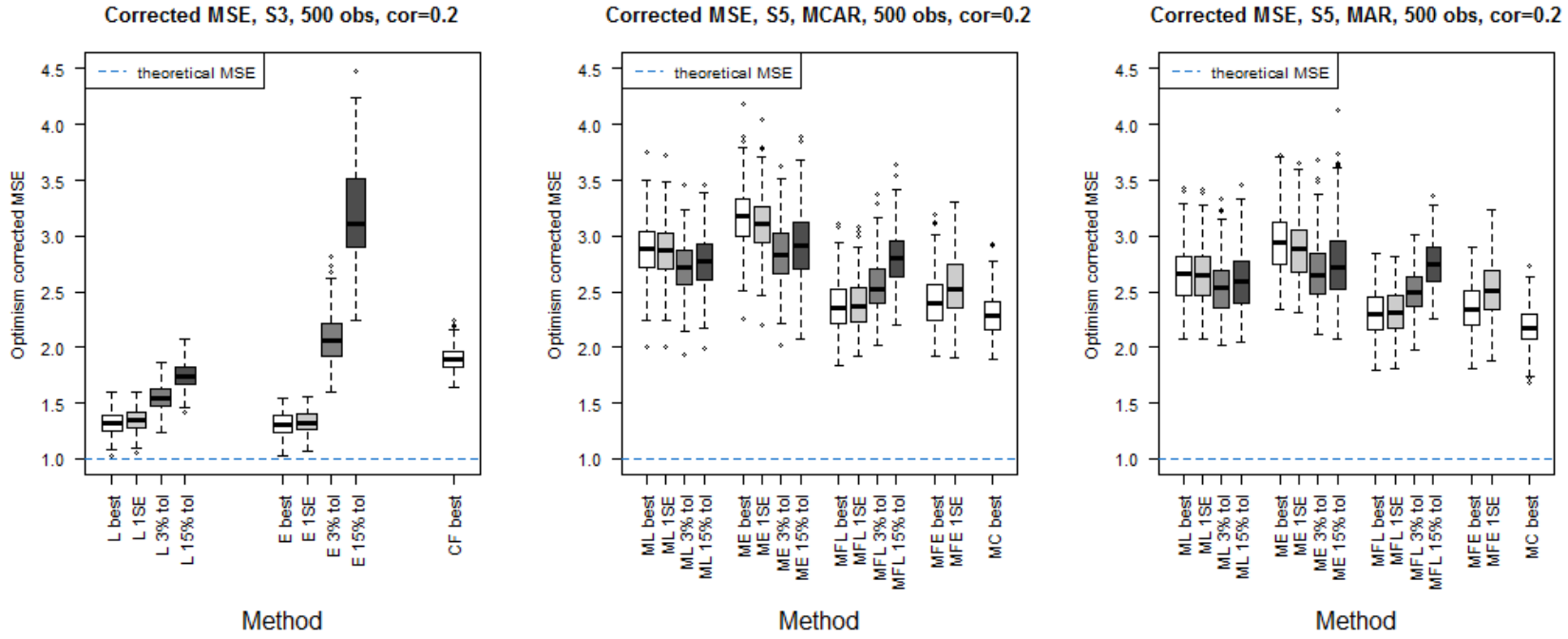


Figure 2.36: **Optimism-corrected MSE** estimates from 5 methods run on 300 simulated **100-covariate** datasets (**correlation 0.8**) with **500 observations** for scenarios S3 (assumption of moderation, without missing data) and S5 (assumption of moderation, missing data also in the outcome). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME), MissForest-Lasso (MFL), MissForest-Elasticnet (MFE), MissForest-Conditional RF (MC). ML, ME, MFL and MFE estimated MSEs are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S3 (first plot from the left), the Lasso (L), Elasticnet (E) and Conditional RF (CF) corrected MSEs are shown.

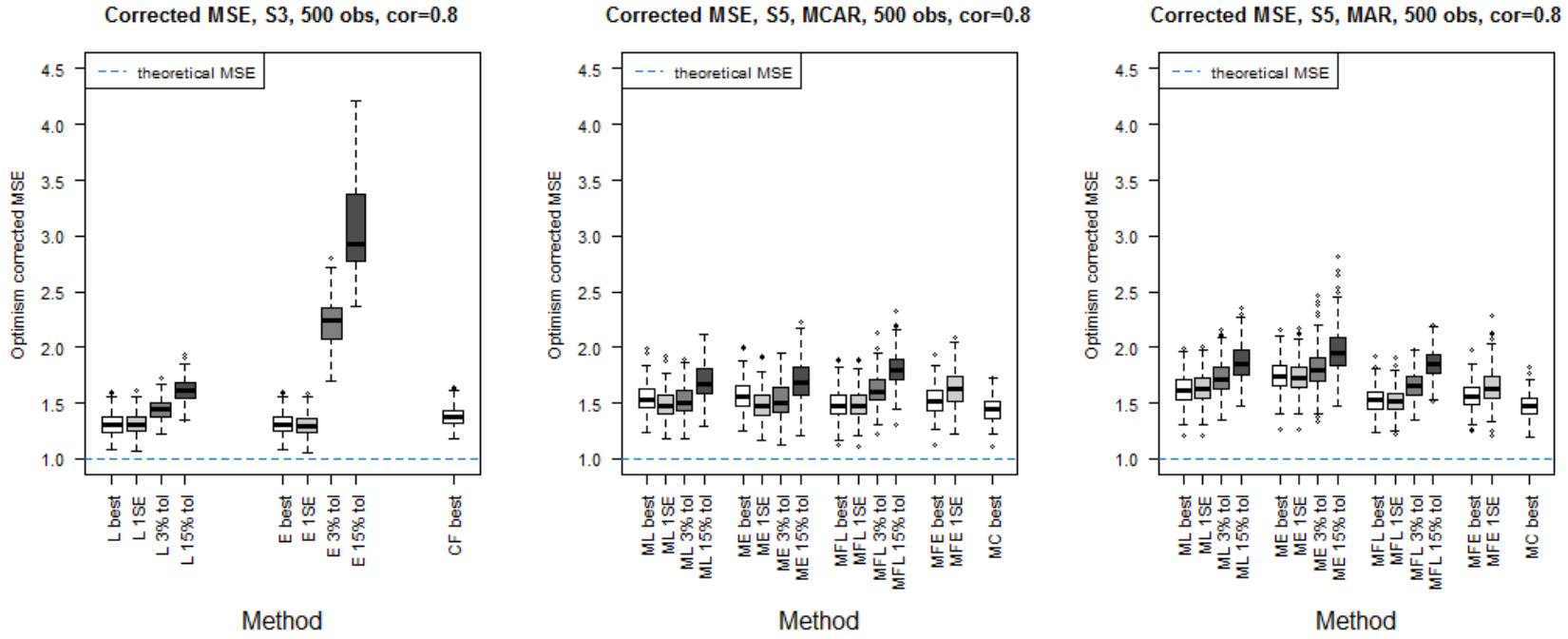


Figure 2.37: **Calibration slope** $\beta_{LP}$ estimates for 5 methods run on 300 simulated **100-covariate** datasets (**correlation = 0.2**) with **500 observations** for scenarios **S3** (assumption of moderation, without missing data) and **S5** (assumption of moderation, missing data also in the outcome). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME), Missforest-Lasso (MFL), MissForest-Elasticnet (MFE) and MissForest-Conditional RF (MC). ML, ME, MFL and MFE estimated calibration slopes are shown for the best $\lambda$ selection as well as for three tolerance models (when available): one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S3 (first plot from the left), the Lasso (L), Elasticnet (E) and Conditional RF (CF) calibration slopes are shown.

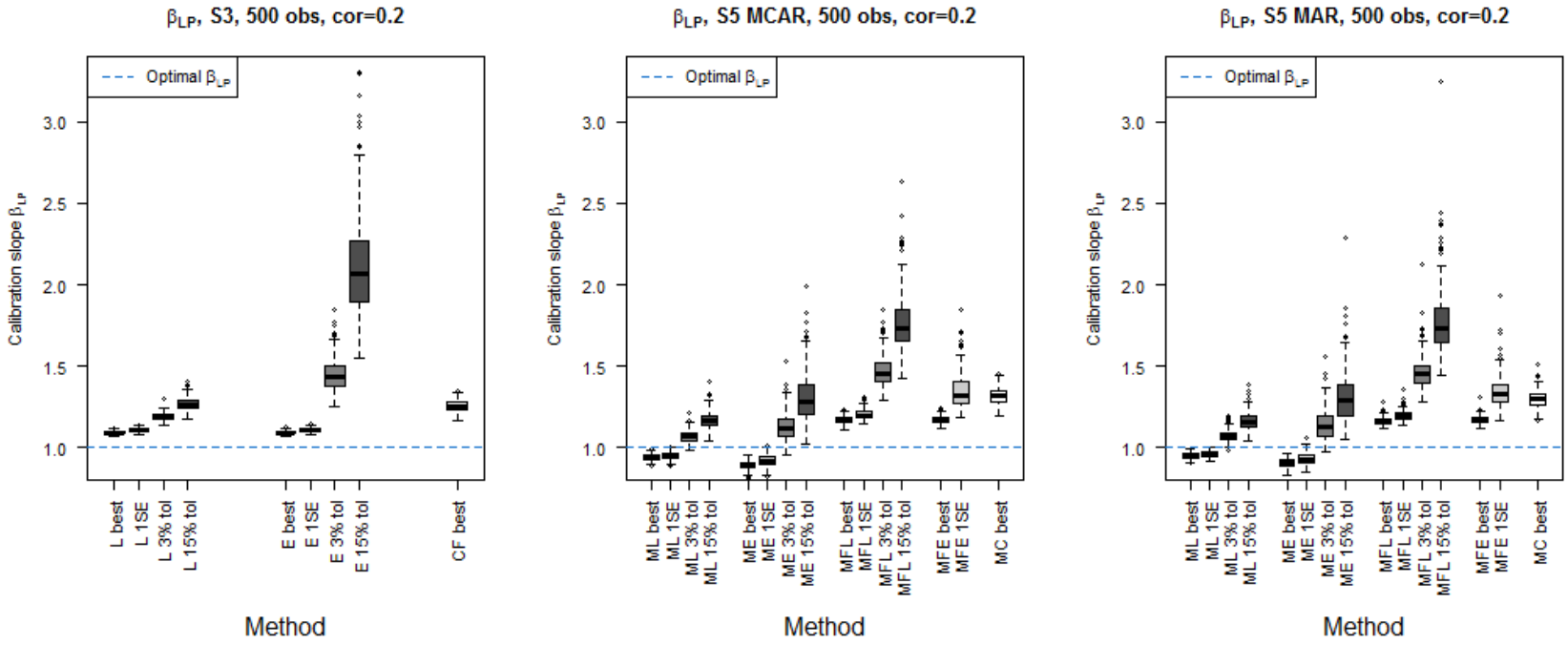


Figure 2.38: **Calibration slope** $\beta_{LP}$ estimates for 5 methods run on 300 simulated **100-covariate** datasets (**correlation = 0.8**) with **500 observations** for scenarios **S3** (assumption of moderation, without missing data) and **S5** (assumption of moderation, missing data also in the outcome). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME), Missforest-Lasso (MFL), MissForest-Elasticnet (MFE) and MissForest-Conditional RF (MC). ML, ME, MFL and MFE estimated calibration slopes are shown for the best $\lambda$ selection as well as for three tolerance models (when available): one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S3 (first plot from the left), the Lasso (L), Elasticnet (E) and Conditional RF (CF) calibration slopes are shown.

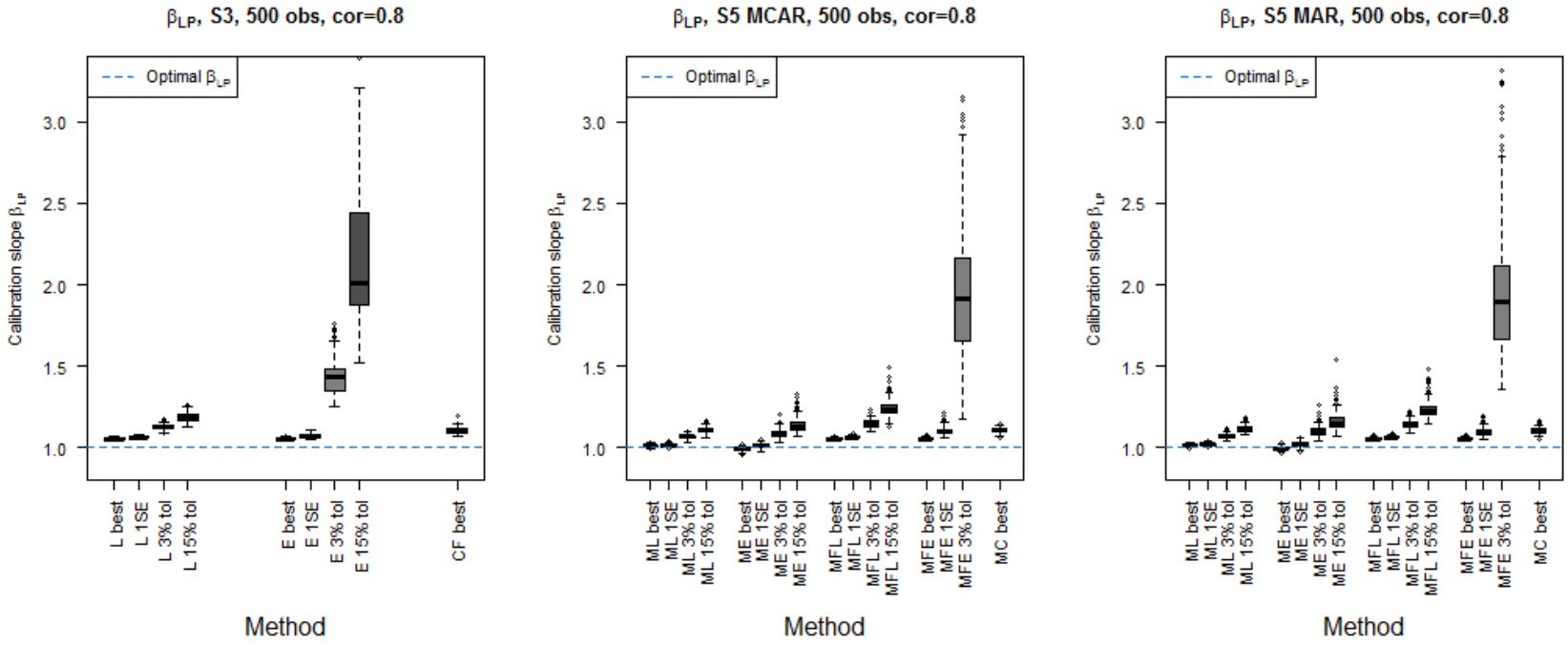


Figure 2.39: Average **internal and external MSE optimism** estimates with 2.5th and 97.5th percentiles for 5 methods run on 300 simulated **100-covariate** datasets (**correlation=0.2**) with **500 observations** for scenarios **S3** (assumption of moderation, without missing data) and **S5** (assumption of moderation, with missing data also in the outcome). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME), MissForest-Lasso (MFL), MissForest-Elasticnet (MFE) and MissForest-Conditional RF (MC). ML, ME, MFL and MFE estimated internal and external MSE optimism are shown for the best $\lambda$ selection as well as for three tolerance models (when available): one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S3 (first plot from the left), the Lasso (L), Elasticnet (E) and Conditional Random Forest (CF) optimism estimates are shown.

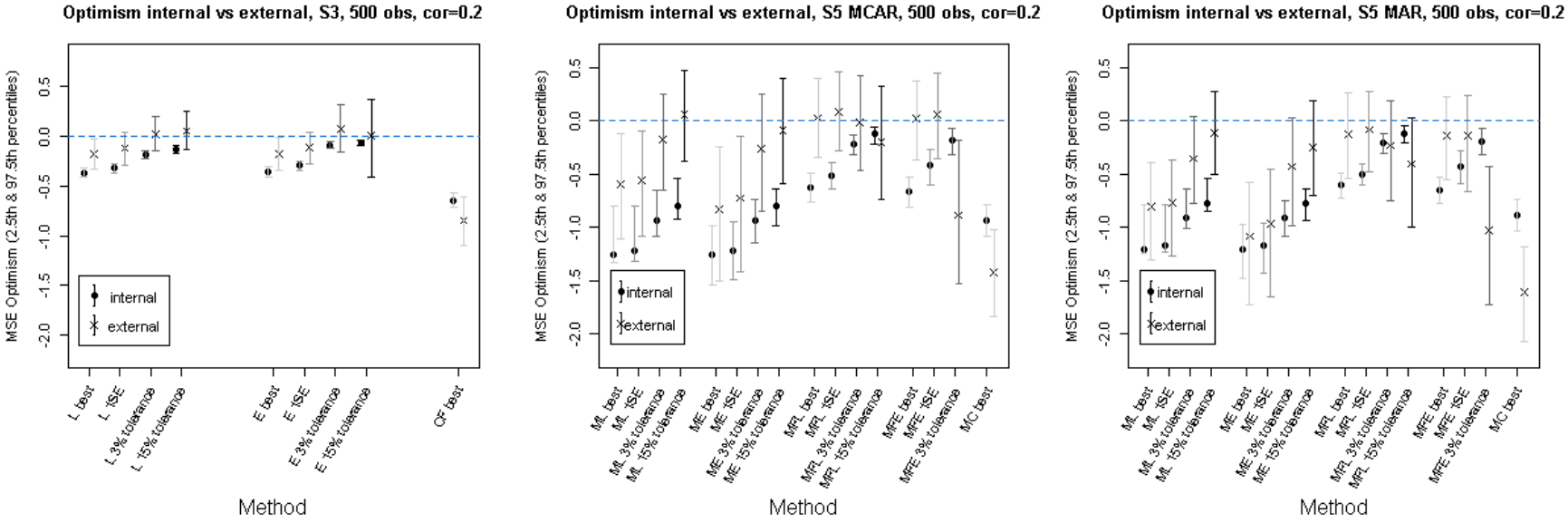


Figure 2.40: Average **internal and external MSE optimism** estimates with 2.5th and 97.5th percentiles for 5 methods run on 300 simulated **100-covariate** datasets (**correlation=0.8**) with **500 observations** for scenarios **S3** (assumption of moderation, without missing data) and **S5** (assumption of moderation, with missing data also in the outcome). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME), MissForest-Lasso (MFL), MissForest-Elasticnet (MFE) and MissForest-Conditional RF (MC). ML, ME, MFL and MFE estimated internal and external MSE optimism are shown for the best $\lambda$ selection as well as for three tolerance models (when available): one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S3 (first plot from the left), the Lasso (L), Elasticnet (E) and Conditional Random Forest (CF) optimism estimates are shown.

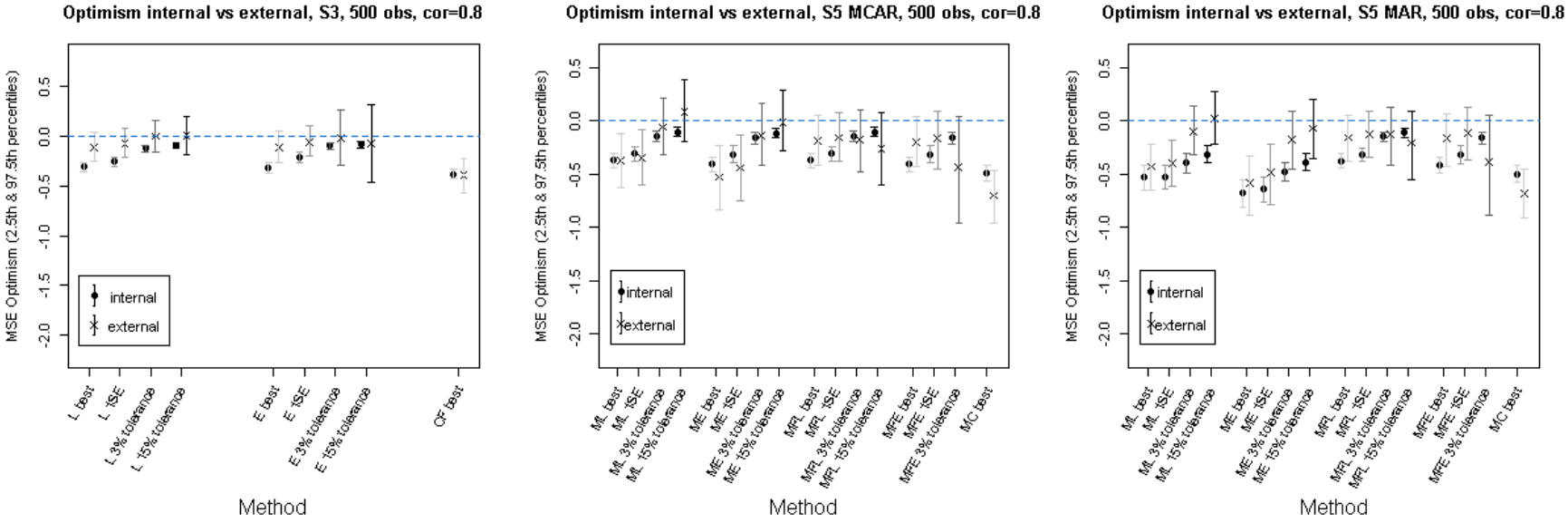


Figure 2.41: Average percentage of **true predictors (TP) selected among the actual TP** (SEN) estimates with 2.5th and 97.5th percentiles from 4 methods run on 300 simulated **100-covariate** datasets (**correlation=0.2**) with **500 observations** for scenarios **S3** (assumption of moderation, without missing data) and **S5** (assumption of moderation, missing data also in the outcome). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME), MissForest-Lasso (MFL) and MissForest-Elasticnet. ML, ME, MFL and MFE estimated percentages of TP selected among the actual TP variables are shown for the best $\lambda$ selection as well as for three tolerance models (when available): one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S3 (first plot from the left), the Lasso (L) and the Elasticnet (E) estimates are shown.

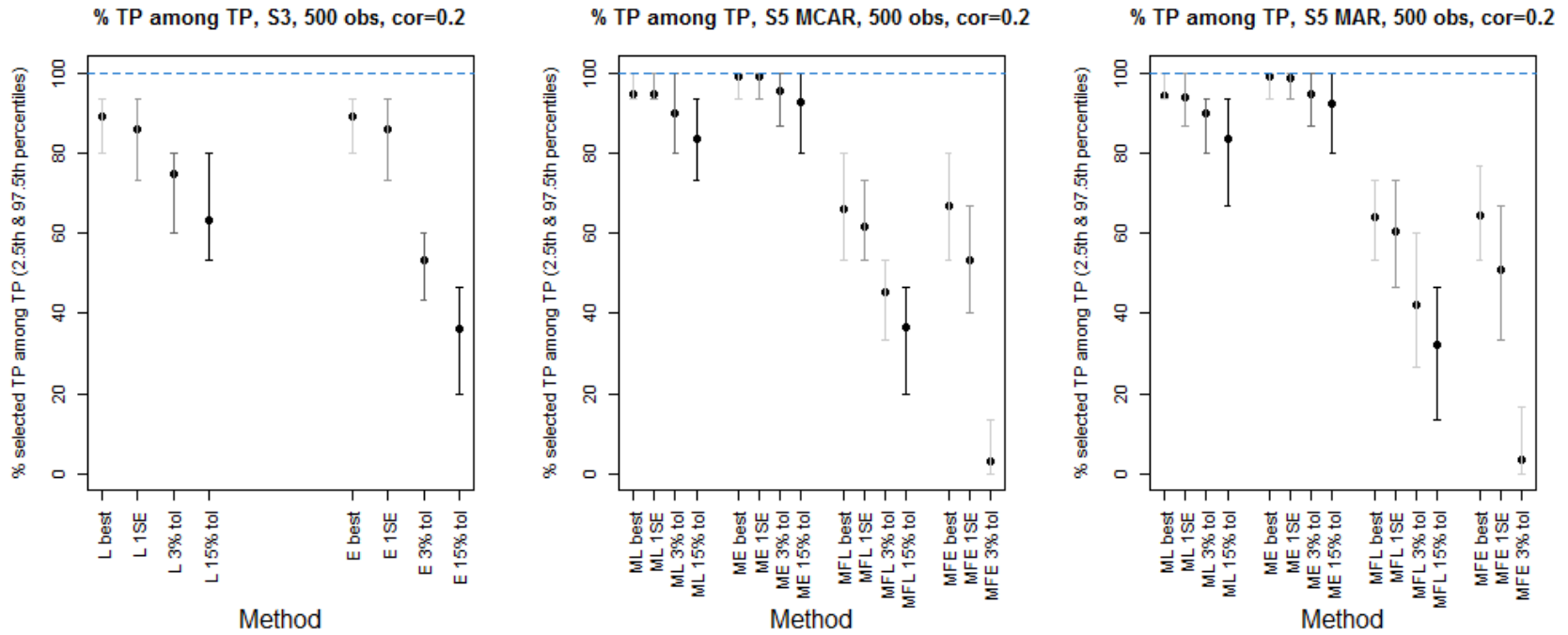


Figure 2.42: Average percentage of **true predictors (TP) selected among the actual TP** (SEN) estimates with 2.5th and 97.5th percentiles from 4 methods run on 300 simulated **100-covariate** datasets (**correlation=0.8**) with **500 observations** for scenarios **S3** (assumption of moderation, without missing data) and **S5** (assumption of moderation, missing data also in the outcome). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME), MissForest-Lasso (MFL) and MissForest-Elasticnet. ML, ME, MFL and MFE estimated percentages of TP selected among the actual TP variables are shown for the best $\lambda$ selection as well as for three tolerance models (when available): one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S3 (first plot from the left), the Lasso (L) and the Elasticnet (E) estimates are shown.

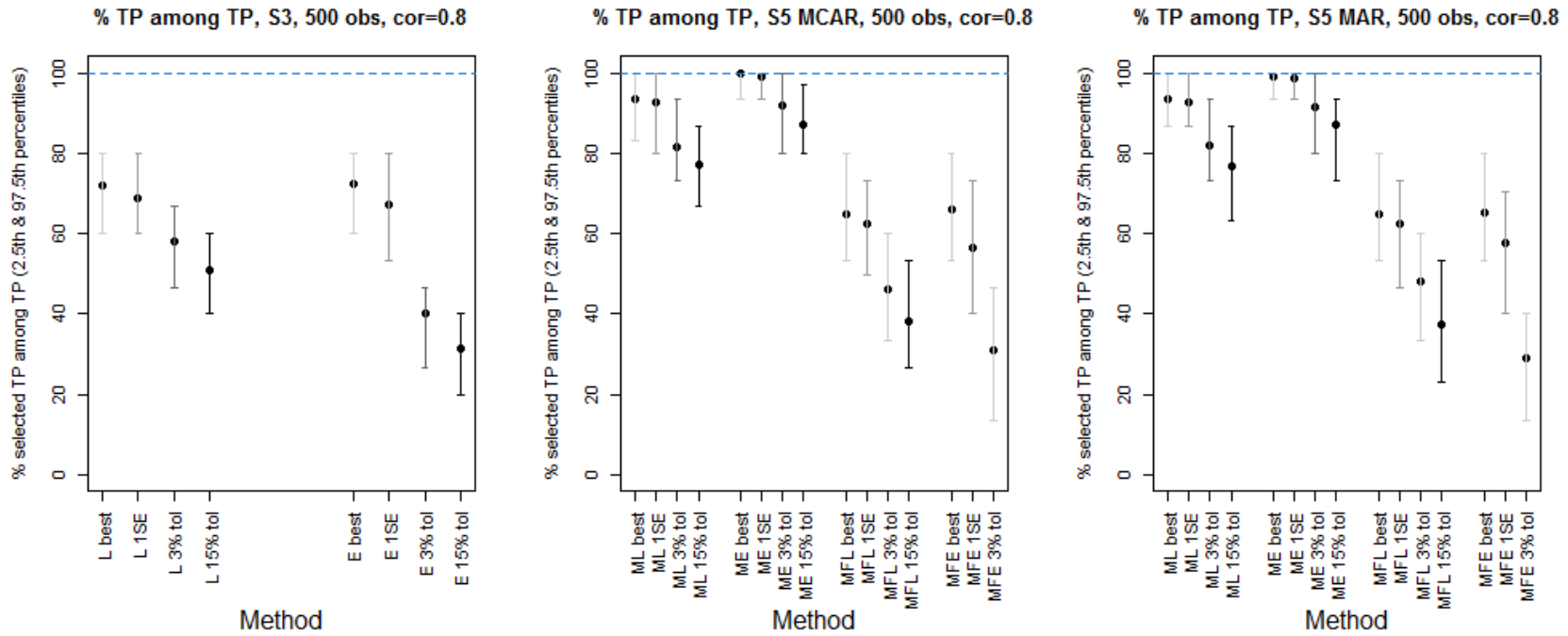


Figure 2.43: Average percentage of **true predictors (TP) among the selected variables** (PPV) estimates with 2.5th and 97.5th percentiles from 3 methods run on 300 simulated **100-covariate** datasets (**correlation=0.2**) with **500 observations** for scenarios **S3** (assumption of moderation, without missing data) and **S5** (assumption of moderation, missing data also in the outcome). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME), MissForest-Lasso (MFL) and MissForest-Elasticnet (MFE). ML, ME and MF estimated percentages of TP among the selected variables are shown for the best $\lambda$ selection as well as for three tolerance models (when available): one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S3 (first plot from the left), the Lasso (L) and the Elasticnet (E) estimates are shown.

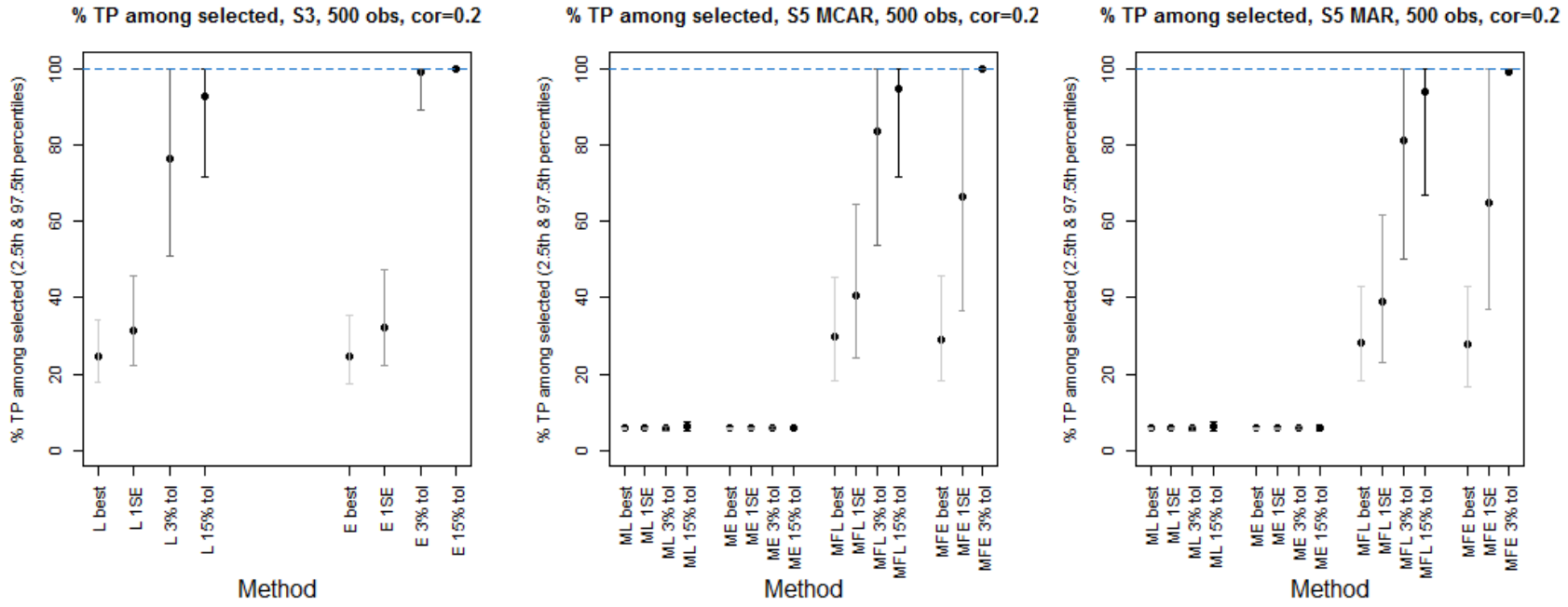
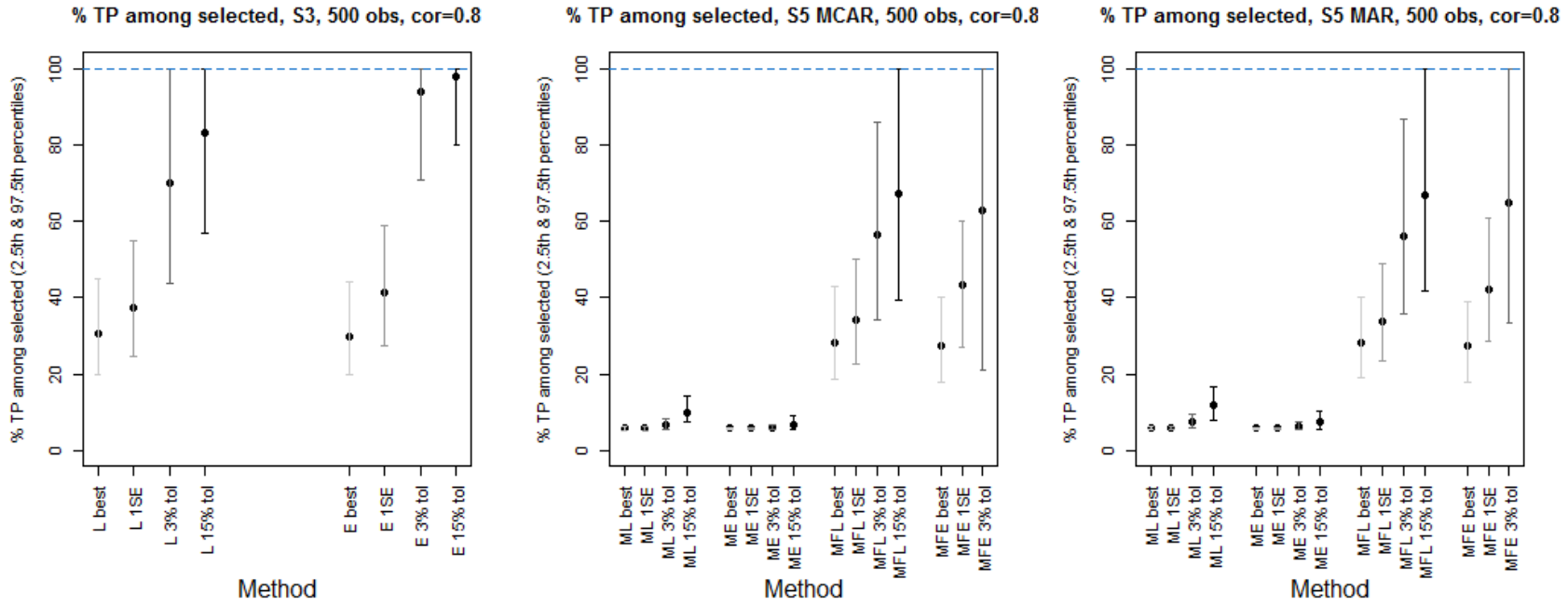


Figure 2.44: Average percentage of **true predictors (TP) among the selected variables** (PPV) estimates with 2.5th and 97.5th percentiles from 3 methods run on 300 simulated **100-covariate** datasets (**correlation=0.8**) with **500 observations** for scenarios **S3** (assumption of moderation, without missing data) and **S5** (assumption of moderation, missing data also in the outcome). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME), MissForest-Lasso (MFL) and MissForest-Elasticnet (MFE). ML, ME and MF estimated percentages of TP among the selected variables are shown for the best $\lambda$ selection as well as for three tolerance models (when available): one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S3 (first plot from the left), the Lasso (L) and the Elasticnet (E) estimates are shown.



### 2.3.3 Selection of moderators

In scenarios S3, S4, S5 and S6 for the 20-covariate study and in the 100-covariate study (scenarios S3 and S5), the selection of true interaction terms identifying moderators was generally as good as or slightly poorer than the selection of the overall true predictors in terms of positive predictive value for the models Mice-Lasso, Mice-Elasticnet, MissForest-Lasso and MissForest-Elasticnet (PPV, see Tables both in the Appendix and here below A.17 for the best models, A.18 for the 1 SE models, A.19 for the 15% tolerance models and 2.36 for the 3% tolerance models). Only MissForest-Lasso 3% and 15% models and MissForest-Elasticnet 1SE and 3% models in the 100-covariate study showed better PPV for moderators especially in the high correlated variables scenario. The discrepancy between predictors and moderators PPV results was largest for the best models and reduced as the tolerance level for the tuning parameter increased.

The sensitivity (SEN) of selection for the moderators was always slighly better than the SEN for the overall predictors in the 20-covariate scenario. In the 100-covariate scenario the SEN for the overall predictors was higher than the SEN for moderators apart from the SEN for the methods Mice-Lasso and Mice-Elasticnet, which however selected all variables without distinguishing between true and false predictors.

The false positive rate (FPR) of selection for the moderators was lower (better) than the FPR for the overall predictors across all scenarios with missing data and for all levels of penalty tolerance.

An acceptable to good moderator selection performance according to our subjective criteria was only achieved in the 20-covariate scenario by the methods in the tolerance models for the complete data scenarios, by MissForest-Lasso for the missing data scenarios in the 3% models (together with Mice-Elasticnet only for 1000 observation datasets) and by MissForest-Lasso and Mice-Lasso in the 15% tolerance models (see Figures 2.36 and A.19 in the Appendix). In the 100-covariate scenario, where the models included 119 interaction terms, of which only 5 terms were true predictors, the selection was generally poor.

Table 2.36: Comparison of average sensitivity (SEN), false positive rate (FPR) and positive predictive value (PPV) of selection for the predictors (P) and for the moderators (M) for the 3% tolerance models in the simulation study. Average SEN, FPR and PPV are given in percentages with corresponding SD.

| | **Selection of predictors and moderators in the 3% tolerance models** | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Scenario 3** | | **Scenario 4: complete outcome** | | | | **Scenario 5: incomplete outcome (20%)** | | | | **Scenario 6: interactions in imputation model** | | | |
| Data | Complete | | MCAR | | MAR | | MCAR | | MAR | | MCAR | | MAR | |
| **20-covariate** | N=250 | N=1000 | N=250 | N=1000 | N=250 | N=1000 | N=250 | N=1000 | N=250 | N=1000 | N=250 | N=1000 | N=250 | N=1000 |
| **Mice-Lasso** | | | | | | | | | | | | | | |
| SEN of P (SD) | 90.5 (4.2) | 99.9 (1.3) | 92.8 (3.4) | 93.5 (1.3) | 93.3 (3.3) | 94.1 (2.2) | 93.7 (3.4) | 93.9 (1.9) | 95.2 (4.1) | 96.8 (3.3) | 94.8 (3.6) | 93.7 (1.6) | 94.2 (3.6) | 93.7 (1.6) |
| SEN of M (SD) | 95.7 (9.5) | 99.8 (2.5) | 98.5 (6.0) | 99.9 (1.4) | 99.2 (4.5) | 100 (0) | 98.6 (5.8) | 100 (0) | 99.3 (4.23) | 100 (0) | 99.6 (3.2) | 100 (0) | 98.8 (5.5) | 100 (0) |
| FPR of P (SD) | 16.6 (7.9) | 2.1 (5) | 52.6 (12.8) | 31.1 (8.3) | 50.7 (12.2) | 32.3 (8.3) | 67.4 (11.8) | 42.5 (9.8) | 67.8 (12.2) | 47 (9.6) | 77.5 (18.4) | 38.2 (11.5) | 69.2 (16.8) | 36.9 (10.5) |
| FPR of M (SD) | 16.7 (9.3) | 9.8 (6.8) | 45.9 (14.1) | 28.6 (10.0) | 44.6 (13.4) | 31.8 (9.6) | 58.1 (14.0) | 37.5 (11.2) | 58.6 (14.2) | 43.0 (10.6) | 74.8 (20.1) | 36.3 (13.0) | 65.9 (18.3) | 35.7 (11.7) |
| PPV of P (SD) | 78.2 (8.3) | 98.2 (6.6) | 53.1 (6.2) | 65.8 (6.2) | 54.2 (6.3) | 65.1 (5.9) | 46.9 (4.6) | 58.6 (5.6) | 47.2 (4.9) | 56.7 (5.1) | 44.2 (6.2) | 61.4 (7.1) | 46.8 (6.4) | 62.1 (6.6) |
| PPV of M (SD) | 63.2 (14.8) | 75.5 (14.2) | 37.9 (8.3) | 50.0 (9.6) | 38.6 (8.2) | 46.9 (8.1) | 32.1 (6.0) | 42.9 (7.8) | 32.1 (6.2) | 39.2 (6.2) | 27.4 (6.4) | 44.2 (9.5) | 29.8 (6.6) | 44.3 (8.4) |
| **Mice-Elasticnet** | | | | | | | | | | | | | | |
| SEN of P (SD) | 76.8 (14.8) | 81.6 (5) | 92.8 (5.2) | 84.9 (5.6) | 93.4 (5.5) | 86.5 (5.2) | 94 (4.7) | 86.4 (6.5) | 95.8 (4.8) | 90 (5.7) | | | | |
| SEN of M (SD) | 85.6 (17.7) | 95.7 (10.1) | 98.1 (7.8) | 95.3 (10.8) | 98.3 (6.6) | 97.4 (7.9) | 98.5 (6.3) | 93.8 (12.4) | 99.3 (4.3) | 98.3 (6.3) | | | | |
| FPR of P (SD) | 11.2 (15) | 1.5 (2.3) | 59.6 (23.9) | 8.6 (10.3) | 60.8 (24.1) | 10.6 (11) | 72.7 (20.1) | 13.9 (12.6) | 74.8 (21.2) | 23.8 (16.9) | | | | |
| FPR of M (SD) | 12.6 (15.5) | 2.4 (3.7) | 57.2 (23.4) | 9.1 (10.9) | 58.0 (23.7) | 11.2 (11.8) | 68.5 (21.6) | 13.7 (12.4) | 70.4 (22.2) | 25.2 (16.2) | | | | |
| PPV of P (SD) | 85.8 (14) | 97.3 (4.1) | 51.8 (11.6) | 88 (10.3) | 51.6 (12) | 85.6 (10.6) | 45.9 (8) | 81.9 (11.4) | 45.9 (8.6) | 73.1 (12) | | | | |
| PPV of M (SD) | 74.2 (21.7) | 92.6 (11.0) | 34.5 (11.8) | 79.1 (18.0) | 34.6 (13.2) | 75.3 (18.2) | 29.6 (8.9) | 70.2 (17.9) | 29.4 (9.5) | 55.7 (15.5) | | | | |
| **MissForest-Lasso** | | | | | | | | | | | | | | |
| SEN of P (SD) | 90.5 (4.2) | 99.9 (1.3) | 86.3 (6.2) | 92.8 (1.8) | 87.5 (5.7) | 92.9 (2) | 82.1 (7.6) | 92.2 (2.8) | 85.4 (6.4) | 91.8 (3) | | | | |
| SEN of M (SD) | 95.7 (9.5) | 99.8 (2.5) | 89.1 (14.3) | 99.4 (3.8) | 90.8 (13.6) | 99.3 (4.0) | 84.6 (15.9) | 98.9 (5.1) | 92.3 (12.6) | 99.7 (2.9) | | | | |
| FPR of P (SD) | 16.6 (7.9) | 2.1 (5) | 16.8 (8) | 9.2 (5.3) | 17.4 (8.3) | 11 (5.9) | 16.4 (7.8) | 10.8 (5.4) | 20.5 (8.8) | 15.2 (7) | | | | |
| FPR of M (SD) | 16.7 (9.3) | 9.8 (6.8) | 15.7 (8.7) | 10.5 (7.0) | 15.9 (9.4) | 11.5 (7.0) | 15.7 (8.8) | 11.4 (6.9) | 19.0 (9.8) | 16.5 (8.0) | | | | |
| PPV of P (SD) | 78.2 (8.3) | 98.2 (6.6) | 77.2 (8.6) | 86.8 (6.7) | 76.9 (8.6) | 84.7 (7) | 76.8 (8.8) | 84.8 (6.8) | 73.3 (8.6) | 79.8 (7.4) | | | | |
| PPV of M (SD) | 63.2 (14.8) | 75.5 (14.2) | 62.7 (14.9) | 74.2 (13.8) | 63.5 (15.4) | 72.0 (13.1) | 61.9 (16.2) | 72.3 (13.6) | 59.2 (14.3) | 63.8 (11.9) | | | | |
| **100-covariate** (N=500) | $\rho = 0.2$ | $\rho = 0.8$ | | | | | $\rho = 0.2$ | $\rho = 0.8$ | $\rho = 0.2$ | $\rho = 0.8$ | | | | |
| **Mice-Lasso** | | | | | | | | | | | | | | |
| SEN of P (SD) | 74.8 (6.2) | 57.9 (5.7) | | | | | 90 (4.7) | 81.5 (5.2) | 89.7 (4.7) | 82 (5.4) | | | | |
| SEN of M (SD) | 67.1 (13.4) | 53.5 (9.5) | | | | | 89.5 (12.3) | 65.5 (13.6) | 88.9 (13.4) | 66.7 (14.6) | | | | |
| FPR of P (SD) | 2.1 (1.3) | 2.2 (1.2) | | | | | 94.7 (2.6) | 74.1 (6.9) | 94.3 (2.7) | 68.1 (8.5) | | | | |
| FPR of M (SD) | 1.1 (1.3) | 0.7 (0.9) | | | | | 89.2 (4.8) | 58.5 (9.2) | 88.4 (4.8) | 53.1 (10) | | | | |
| PPV of P (SD) | 76.4 (13) | 70.1 (13.6) | | | | | 5.8 (0.3) | 6.7 (0.7) | 5.8 (0.3) | 7.3 (0.9) | | | | |
| PPV of M (SD) | 78.2 (20.1) | 82.3 (20.1) | | | | | 4 (0.6) | 4.6 (1.1) | 4.1 (0.6) | 5.2 (1.3) | | | | |
| **Mice-Elasticnet** | | | | | | | | | | | | | | |
| SEN of P (SD) | 53.1 (4.9) | 40.3 (5.8) | | | | | 95.3 (3.9) | 92 (5.2) | 94.7 (3.5) | 91.4 (5.3) | | | | |
| SEN of M (SD) | 43.9 (8.3) | 46.7 (13.8) | | | | | 98.9 (4.9) | 91.3 (12.3) | 98.7 (5) | 90.9 (11.8) | | | | |
| FPR of P (SD) | 0.5 (0.1) | 0.6 (0.3) | | | | | 99.7 (1.7) | 93.7 (5.5) | 99.5 (1.7) | 90.5 (7.5) | | | | |
| FPR of M (SD) | 0 (0.2) | 0 (0.1) | | | | | 98.6 (3) | 88.3 (8.8) | 98.2 (3) | 83.9 (10.8) | | | | |
| PPV of P (SD) | 99 (3.2) | 93.9 (9.7) | | | | | 5.8 (0.2) | 6 (0.4) | 5.8 (0.2) | 6.1 (0.5) | | | | |
| PPV of M (SD) | 99 (5.4) | NA | | | | | 4 (0.2) | 4.2 (0.6) | 4.1 (0.2) | 4.4 (0.7) | | | | |
| **MissForest-Lasso** | | | | | | | | | | | | | | |
| SEN of P (SD) | 74.8 (6.2) | 57.9 (5.7) | | | | | 45.7 (6.4) | 46.5 (7.9) | 42.3 (8.4) | 48.1 (7.4) | | | | |
| SEN of M (SD) | 67.1 (13.4) | 53.5 (9.5) | | | | | 38.7 (10.8) | 48.4 (11.5) | 34.1 (14.7) | 44.8 (12.1) | | | | |
| FPR of P (SD) | 2.1 (1.3) | 2.2 (1.2) | | | | | 1.2 (0.7) | 3 (1.3) | 1.2 (0.8) | 3 (1.3) | | | | |
| FPR of M (SD) | 1.1 (1.3) | 0.7 (0.9) | | | | | 0.5 (0.7) | 1 (1.2) | 0.4 (0.6) | 1.2 (1.4) | | | | |
| PPV of P (SD) | 76.4 (13) | 70.1 (13.6) | | | | | 83.1 (14.1) | 56.2 (12.9) | 81.2 (14.5) | 57 (13.2) | | | | |
| PPV of M (SD) | 78.2 (20.1) | 82.3 (20.1) | | | | | 83.8 (21.3) | 74.4 (23.6) | 83.9 (24.1) | 69.5 (26.1) | | | | |
| **MissForest-Elasticnet** | | | | | | | | | | | | | | |
| SEN of P (SD) | 53.1 (4.9) | 40.3 (5.8) | | | | | 3.2 (4.8) | 31.6 (8.2) | 3.5 (5.4) | 29.3 (7.1) | | | | |
| SEN of M (SD) | 43.9 (8.3) | 46.7 (13.8) | | | | | 0.9 (4.4) | 30.4 (12.3) | 1.6 (5.7) | 26.7 (11) | | | | |
| FPR of P (SD) | 0.5 (0.1) | 0.6 (0.3) | | | | | 0.4 (0) | 2.4 (2.5) | 0.4 (0.1) | 1.6 (1) | | | | |
| FPR of M (SD) | 0 (0.2) | 0 (0.1) | | | | | 0 (0) | 0.2 (0.9) | 0 (0.1) | 0.1 (0.4) | | | | |
| PPV of P (SD) | 99 (3.2) | 93.9 (9.7) | | | | | 99.8 (2.3) | 61.4 (22) | 99.1 (5.9) | 65.7 (18.7) | | | | |
| PPV of M (SD) | 99 (5.4) | NA | | | | | 95.8 (14.4) | 93.8 (16.5) | 95.7 (14.4) | 95.1 (14.1) | | | | |

## 2.4 Summary of results

**Summary for 20-covariate data simulation study**

MICE-Lasso, MICE-Elasticnet, MissForest-Lasso and MissForest-RF were compared in terms of prediction accuracy and variable selection in the 20-covariate simulation study (see Subsection 2.2.1).

Overall MissForest-Lasso performed better in accuracy and variable selection compared to the other methods in both 250 and 1000 observations dataset cases across scenarios with missing data.

The models showing at the same time an acceptable prediction accuracy and good variable selection were the 1 SE and 3% tolerance Lasso models in all scenarios without missing data and the 1 SE and 3% tolerance models of MissForest-Lasso in the missing data scenarios where there was no assumption of moderation and the outcome was complete (for 3% tolerance models estimates see the summary Table 2.37 and the summary Figure 2.45).

**Prediction accuracy results for 20-covariate data study** The optimism-corrected MSE was in all simulations above the theoretical MSE. The corrected MSEs increased with increasing penalty and with introducing missingness and was higher if missingness was MCAR. The corrected MSE was poor (then not acceptable) for all the 15% tolerance models. The corrected MSE for the best, 1 SE and 3% tolerance models were acceptable (within 30% of the theoretical MSE) for scenarios S1 (no moderation assumption, complete data), S2 (no moderation assumption, missing data, complete outcome) and S3 (moderation assumption and complete data) for MissForest-Lasso and MICE-Lasso. The corrected MSE for the best, 1SE and 3% tolerance MissForest-Lasso models in scenario S5 (moderation assumption, missing data also in the outcome) were also acceptable in the case of MAR data.

In scenario S1, without missing data and in absence of moderator variables among the predictors, Lasso outperformed Elasticnet and RF in prediction accuracy when the sample size was 250 (see Figures 2.8). In scenario S1, the difference between MSE internal and external optimism was the least for RF (see Figure 2.14), indicating that its larger variance performance is also what will be obtained on new data.

In scenario S2, when missing data were present with a complete outcome, the methods performance overall slightly decreased: the regularised regression methods combined with MICE or MissForest best and 1SE tolerance models had similar accuracy, which was superior to MissForest-RF that seemed to underfit the data.

When interactions were added to the linear predictor in scenario S3 without missing data, Lasso had again the best performance while Elasticnet tolerance models and RF were underfitting the data (see Figures 2.9). In scenario S4 with missing data, the accuracy was moderately inferior alike for scenario S2: the best and 1SE tolerance models of MICE-Lasso and MissForest-Lasso had similar accuracy and their result was the best among the methods.

All methods performed their worst in scenario S5 when missing data were in the outcome, apart from MissForest-Lasso that maintained a performance similar to scenario S4 with complete outcome (see Figure 2.10). However, the difference between internal and external MSE optimism revealed that this estimated discrimination was too optimistic for MissForest-Lasso (see Figure 2.16).

In scenario S6, including interactions in the MICE imputation model introduced more noise in the analysis compared to scenario S4 (see Figure 2.2).

Calibration performance for each method was constant across scenarios and MICE-Lasso and MissForest-Lasso best and 1SE models had the best average calibration slope estimates (see Figures 2.11, 2.12, 2.13 and 2.3).

All the accuracy results of the 1000 observation datasets analyses were very much improved compared to the smaller sample size (see Figures A.5, A.6, A.7, A.8, A.9, A.10, A.11, A.12, A.13 in Appendix).

**Variable selection results for 20-covariate data study** In all scenarios with datasets of 250 observations, all methods selected more than 70% of the true predictors (TP) with higher percentages in the low tolerance models apart from Elasticnet and MICE-Elasticnet, which had poor sensitivity of selection (SEN) in the 15% tolerance models. Also false-positive predictors (FP) were selected, increasing in number with introducing missing data, with the highest (and very poor) false positive rate of selection (FPR) for the best models of MICE-Lasso and MICE-Elasticnet. However, with increasing penalty the number of FP decreased, leading to better positive predictive values (PPV) and good variable selection performance for the 1 SE, 3% and 15% tolerance models of Lasso, Elasticnet and MissForest-Lasso in all scenarios (apart from the 1 SE tolerance model in scenario S5 with incomplete outcome for MissForest-Lasso, which had only an acceptable performance).

In scenarios S1 (without assumption of moderation) and S3 (with assumption of moderation) in absence of missing data with 250 observations, Lasso and Elasticnet selected all the TPs almost always in the best and 1SE tolerance models (see Figures 2.17 and 2.18), but included also many noise predictors (see Figures 2.20 and 2.21). Rarely only the TPs were selected all

together (see Figure 2.23 and tables 2.3, 2.5, 2.17, 2.23, for S1 and 2.7, 2.8, 2.19 and 2.26 for S3) and slightly more often all the TP were chosen at the same time apart from one TP (see Figures 2.24 for S1 and 2.25 for S3). Shrinkage in Elasticnet tolerance models was stronger compared to Lasso given the same level of penalty tolerance. Instead RF had individual TPs included in the 10 most important variables between only 20% and 100% of the times (see Figure 2.27) and had the 10 TPs included in the top 10 important variables in up to 2.3% of the times and 9 of the TP in up to 37.0% of simulations (see table 2.23) for S1, and never for S3.

MissForest-Lasso best models variable selection performance stands out in scenarios S2 and S4 (with missing data) because it is the only one comparable to Lasso's in scenario S1 (see Figures 2.29, 2.28, 2.17 and 2.20 for S2 and figures 2.33, 2.31, 2.18 and 2.21 for S4). On the contrary, MICE-Lasso and MICE-Elasticnet selected all the variables almost always without distinction between TPs and FPs. As opposed to this, MissForest-RF variable selection was consistent with RF performance in scenario S1.

Variable selection performance was slightly negatively affected by the fact that the outcome was not complete in scenario S5: all methods selected variables with similar pattern as in scenario S4 (see Figures 2.19, 2.22, 2.26, 2.34 and 2.32).

The moderator selection performance (see Subsection 2.3.3) was overall slightly inferior than the general variable selection performance.

When the sample size was 1000, again all results ameliorated by maintaining the same patterns between methods (see Figures A.14, A.15,A.16, A.17, A.18, A.19, A.20, A.21, A.22, A.23,A.24, A.25, A.26, A.27, A.28, A.29, A.30, A.31, A.2, A.3 in the Appendix).

In conclusion, MissForest-Lasso performed best overall.

**Summary for 100-covariate data simulation study**

This simulation study based on 100-covariate data assessed variable selection and accuracy performance for five combined methods, i.e. MICE-Lasso, MICE-Elasticnet, MissForest-Lasso, MissForest-Elasticnet and MissForest-Conditional RF, in the scenarios S3 (complete data, assumption of moderation) and S5 (missing data also in the outcome, assumption of moderation) with two sub-scenarios: correlation between covariates being 0.2 or 0.8.

All methods performed poorly in prediction accuracy in presence of missing data, according to my subjective criteria for model used in clinical practice (see Subsection 2.2) and the methods delivering the best discrimination were MissForest-Lasso best model and MissForest-Conditional RF. The best variable selection performance in the scenario with missing data was MissForest-Conditional RF in both correlation settings. The 3% tolerance MissForest-Lasso

model showed relatively good variable selection results, which were the closest to MissForest-Conditional RF results when correlation was low.

Among the regularised regression methods, MissForest-Lasso tolerance model results were best in accuracy and variable selection as for the 20-covariate data simulation study (see the summary Table 2.37 and the summary Figure 2.45).

**Prediction accuracy result for 100-covariate data study** In absence of missing data (S3), Lasso outperformed Elasticnet and Conditional RF in prediction accuracy for both correlation sub-scenarios (see Figures 2.35, 2.36, 2.37, 2.38, 2.39 and 2.40) as it happened in the 20-covariate data study (see Subsection 2.3.1).

When missing data were present (S5), MissForest-Lasso and MissForest-Conditional RF were superior to the other methods but still performed relatively poorly according to my definition for clinical use (see Figures 2.35 and 2.36). However, calibration for MissForest-Conditional RF in the low correlation scenario was the poorest among the methods (see Figures 2.37 and 2.38) with MissForest-Lasso coming next, suggesting that recalibration of models is usually necessary. Also the mean MSE internal and external optimism were the largest in absolute values and the farthest from each other for MissForest-Conditional RF in the low correlation scenario (followed by MissForest-Lasso best and 1SE models, however the optimism estimates were smaller for MissForest-Lasso and close to each other for the higher tolerance models, see Figures 2.39 and 2.40).

Overall, the methods accuracy results improved with higher correlation between covariates.

**Variable selection result for 100-covariate data study** In scenario S3, as the vector of true coefficients was sparse (15 non-zero entries vs 234 zeros) and the variables were equally correlated, Lasso and Elasticnet best penalty models chosen variables on average contained more FPs than TPs. Instead, only the models with 3% and higher tolerance penalties retained variables with higher percentage of TPs than FPs in both correlation scenarios (see Figures 2.43 and 2.44).

In scenario S5, MissForest-Conditional RF always included the 15 TPs in the top 15 important variables showing that conditional trees RF variable importance (VI) measure is less biased than the traditional RF VI. The latter would have preferred the continuous variables with larger probability of missingness, instead the conditional VI treats all variable alike. On the other hand, MICE-combined methods selected both TPs and FPs altogether almost all the times (same as MICE methods in the 20-covariate scenario), while MissForest-Lasso tended to select higher

percentages of FPs than TPs with the best penalties and to choose higher percentages of TPs with the tolerance penalties for acceptable PPV (in the same way as Lasso). However, the 3% MissForest-Lasso tolerance model performed reasonable well in the low correlation setting with mean positive predictive value of selection larger than 80% (see Figures 2.41 and 2.42).

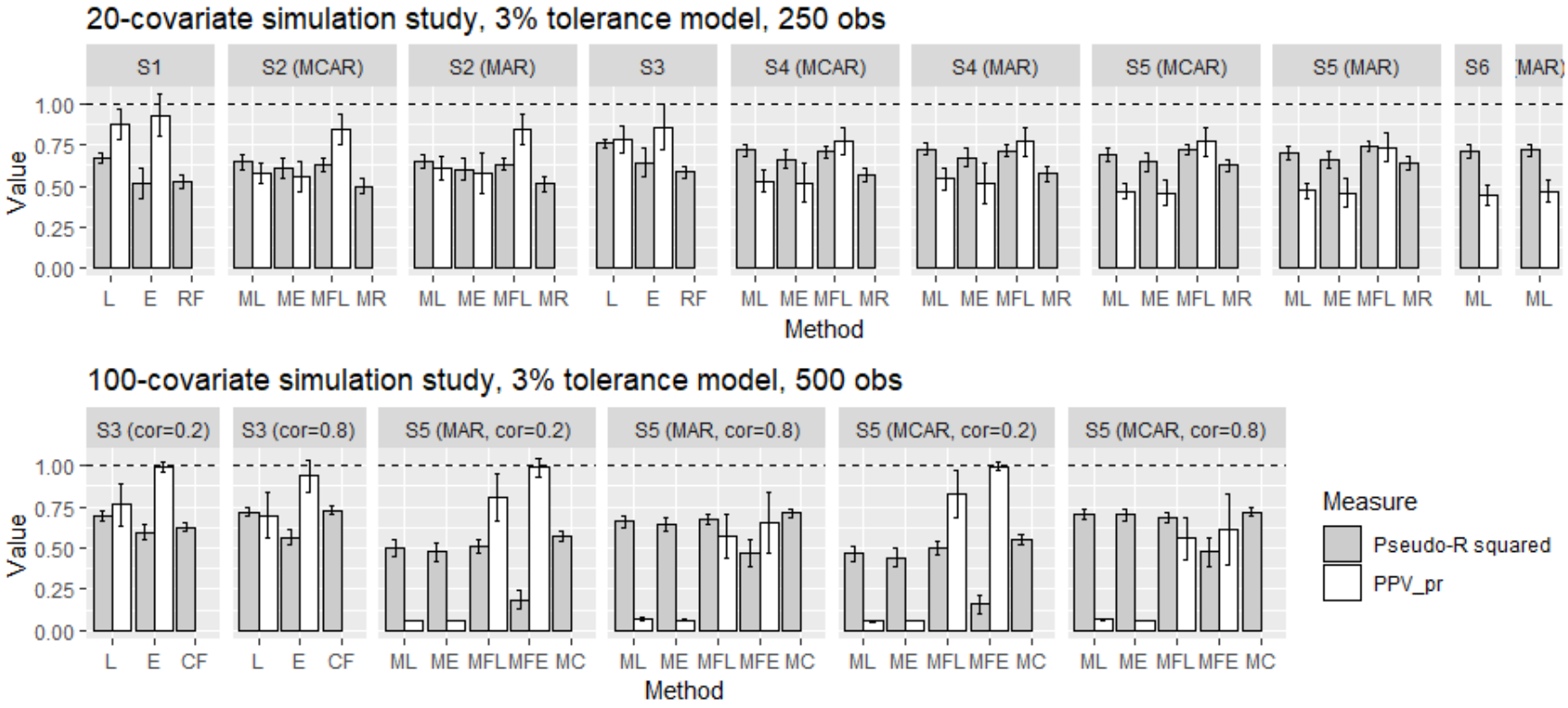
The moderator selection performance was generally poorer than the overall true predictors variable selection in terms of SEN in the low correlation scenarios for MissForest-Lasso and better in the high correlation case. The PPV of moderators was better than the PPV for the overall predictors for MissForest-Lasso (see Subsection 2.3.3).

In general the low correlation scenario had better variable selection results.

Table 2.37: Summary table of results for the simulation study. The average corrected pseudo-$R^2$ and positive predictive value of selection (PPV) for the predictors, showed as a proportion ($PPV_{pr}$ =PPV/100), are presented for all the methods and the main scenarios in the simulation study. The 20-covariate study results are given for the datasets with 250 observations only. Mice-Lasso, Mice-Elasticnet, MissForest-Lasso and MissForest-Elasticnet estimates are shown for the 3% tolerance model only. Scenario 1: complete data, no moderation; Scenario 2: missing data, no moderation, complete outcome; Scenario 3: complete data, moderation; Scenario 4: missing data, moderation, complete outcome; Scenario 5: missing data also in the outcome, moderation; Scenario 6: missing data, moderation, complete outcome, interaction terms in the imputation model. For Scenarios 1 and 3, the Lasso, the Elasticnet, Random Forests and Conditional Random Forests estimates are shown.

| | **Summary table of results for 3% tolerance models** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Data | No moderation | | | Moderation | | | | | | |
| | **Scenario 1** | **Scenario 2** | | **Scenario 3** | **Scenario 4** | | **Scenario 5** | | **Scenario 6** | |
| **20-covariate** (N=250) | Complete | MCAR | MAR | Complete | MCAR | MAR | MCAR | MAR | MCAR | MAR |
| **Mice-Lasso** | | | | | | | | | | |
| pseudo-$R^2$ | 0.67 (0.03) | 0.65 (0.04) | 0.65 (0.04) | 0.76 (0.03) | 0.72 (0.04) | 0.72 (0.03) | 0.69 (0.04) | 0.70 (0.04) | 0.71 (0.04) | 0.72 (0.04) |
| $PPV_{pr}$ | 0.87 (0.09) | 0.58 (0.06) | 0.61 (0.07) | 0.78 (0.08) | 0.53 (0.06) | 0.54 (0.06) | 0.47 (0.05) | 0.47 (0.05) | 0.44 (0.06) | 0.47 (0.06) |
| **Mice-Elasticnet** | | | | | | | | | | |
| pseudo-$R^2$ | 0.52 (0.09) | 0.61 (0.06) | 0.60 (0.07) | 0.64 (0.08) | 0.66 (0.06) | 0.67 (0.06) | 0.64 (0.05) | 0.66 (0.05) | | |
| $PPV_{pr}$ | 0.93 (0.13) | 0.55 (0.09) | 0.58 (0.12) | 0.86 (0.14) | 0.52 (0.12) | 0.52 (0.12) | 0.46 (0.08) | 0.46 (0.09) | | |
| **MissForest-Lasso** | | | | | | | | | | |
| pseudo-$R^2$ | 0.67 (0.03) | 0.63 (0.04) | 0.63 (0.04) | 0.76 (0.03) | 0.71 (0.04) | 0.71 (0.03) | 0.72 (0.03) | 0.74 (0.03) | | |
| $PPV_{pr}$ | 0.87 (0.09) | 0.84 (0.09) | 0.85 (0.09) | 0.78 (0.08) | 0.77 (0.09) | 0.77 (0.09) | 0.77 (0.09) | 0.73 (0.09) | | |
| **MissForest-Random Forests** | | | | | | | | | | |
| pseudo-$R^2$ | 0.52 (0.04) | 0.50 (0.05) | 0.51 (0.05) | 0.58 (0.04) | 0.57 (0.04) | 0.57 (0.04) | 0.63 (0.04) | 0.64 (0.04) | | |
| Data | | | | **Scenario 3** | | **Scenario 5** | | | | |
| **100-covariate** (N=500) | | | | Complete | | MCAR | | MAR | | |
| Correlation between variables | | | | $\rho = 0.2$ | $\rho = 0.8$ | $\rho = 0.2$ | $\rho = 0.8$ | $\rho = 0.2$ | $\rho = 0.8$ | |
| **Mice-Lasso** | | | | | | | | | | |
| pseudo-$R^2$ | | | | 0.63 (0.02) | 0.73 (0.02) | 0.55 (0.03) | 0.72 (0.03) | 0.57 (0.03) | 0.71 (0.03) | |
| $PPV_{pr}$ | | | | 0.99 (0.03) | 0.94 (0.10) | 0.06 (0.00) | 0.06 (0.00) | 0.06 (0.00) | 0.06 (0.00) | |
| **Mice-Elasticnet** | | | | | | | | | | |
| pseudo-$R^2$ | | | | 0.70 (0.03) | 0.72 (0.02) | 0.50 (0.04) | 0.68 (0.03) | 0.51 (0.04) | 0.68 (0.03) | |
| $PPV_{pr}$ | | | | 0.99 (0.03) | 0.94 (0.10) | 0.06 (0.00) | 0.06 (0.00) | 0.06 (0.00) | 0.06 (0.00) | |
| **MissForest-Lasso** | | | | | | | | | | |
| pseudo-$R^2$ | | | | 0.70 (0.03) | 0.72 (0.02) | 0.50 (0.04) | 0.68 (0.03) | 0.51 (0.04) | 0.68 (0.03) | |
| $PPV_{pr}$ | | | | 0.76 (0.13) | 0.70 (0.14) | 0.83 (0.14) | 0.56 (0.13) | 0.81 (0.14) | 0.57 (0.13) | |
| **MissForest-Elasticnet** | | | | | | | | | | |
| pseudo-$R^2$ | | | | 0.59 (0.04) | 0.57 (0.04) | 0.16 (0.05) | 0.48 (0.08) | 0.19 (0.05) | 0.47 (0.08) | |
| $PPV_{pr}$ | | | | 0.99 (0.03) | 0.94 (0.10) | 1.00 (0.02) | 0.61 (0.22) | 0.99 (0.06) | 0.66 (0.19) | |
| **MissForest-Conditional Random Forests** | | | | | | | | | | |
| pseudo-$R^2$ | | | | 0.63 (0.02) | 0.73 (0.02) | 0.55 (0.03) | 0.72 (0.03) | 0.57 (0.03) | 0.71 (0.03) | |

Figure 2.45: Summary figure of results for the simulation study. The average corrected pseudo-$R^2$ and positive predictive value of selection (PPV) for the predictors, showed as a proportion ($PPV_{pr}$ =PPV/100), are presented for all the methods and the main scenarios in the simulation study. The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME), MissForest-Lasso (MFL) and MissForest-Elasticnet (MFE). The 20-covariate study results are given for the datasets with 250 observations only. Mice-Lasso, Mice-Elasticnet, MissForest-Lasso and MissForest-Elasticnet estimates are shown for the 3% tolerance model only. Scenarios abbreviations: S1: complete data, no moderation; S2: missing data, no moderation, complete outcome; S3: complete data, moderation; S4: missing data, moderation, complete outcome; S5: missing data also in the outcome, moderation; S6: missing data, moderation, complete outcome, interaction terms in the imputation model. For S1 and S3, the Lasso (L), the Elasticnet (E), Random Forests (RF) and Conditional Random Forests (CF) estimates are shown.



## 2.5 Discussion and conclusion

In this chapter, I assessed strategies to combine statistical learning prediction modelling techniques and imputation methods in order to obtain unbiased estimates of various predictive performance measures (MSE, MSE optimism and calibration slope) and to perform reliable variable selection. My investigation covered a range of data set characteristics: different sample sizes, missing data scenarios, correlation matrices, types of variables and relationship between predictors and outcome (i.e. assumption of moderation).

In Subsection 2.1.5, I presented ten hypothesis of what I would have expected as results from the simulation study. The following is what I obtained:

- Lasso needed a stronger penalty than the optimal penalty (returning the smallest average MSE) with little loss of prediction accuracy to correct for its model selection inconsistency (high false positive rate, Fan and Lv 2009) and deliver good variable selection when the ratio between number of covariates and sample size was higher, as expected (hypothesis 1). In the 20-covariate dataset complete data scenarios with sample size = 250, the 1 SE tolerance penalty was sufficient to return a high sensitivity of selection ($>$70%) for the Lasso and reach good variable selection performance with a prediction accuracy comparable to the optimum. This is consistent with the literature (Breiman, JH Friedman, et al. 1984 and Hastie, Tibshirani, and Friedman 2008). Instead, in the 100-covariate dataset study (sample size = 500) complete data case, a stronger penalty was needed (3%) to achieve a high positive predictive value ($>$70%) for good variable selection. The strongest assessed 15% tolerance typically showed poor prediction accuracy. However, selected variables of the 15% tolerance model were in most scenarios true predictors. Therefore, the 15% tolerance model may be considered if the selection of a parsimonious set of strongest true predictors is the key research question or to identify such in secondary analyses.

- Elasticnet did not outperform Lasso when predictors were highly correlated, contrarily to what I expected (hypothesis 2). Elasticnet had similar best model performance to the Lasso in both variable selection and prediction accuracy. However, using stronger penalties resulted in poorer performance of Elasticnet compared to Lasso. This unexpected pattern may happen because, in the high correlation case (0.8), not only were the true predictors (TP) correlated between each other, but they were also equally highly correlated with false positives variables (FP). A similar pattern was described by Lu and Petkova (2014). Also, when the correlation matrix was sparse with mixed high and low

correlations (20-covariate case), the presence of some high correlation between FP and TP (see equation (2.19)) negatively affected the variable selection performance of Elasticnet.

- Random Forests did not perform best in prediction accuracy, contrarily to what I expected (hypothesis 3). Instead, it underfitted the data in all complete data scenarios giving the worst predictive performance. This was more pronounced when the sample size was small (250), there was sparse and low correlation between TP and sparse and low to moderate correlation between TP and FP in the 20-covariate simulation study. Also Conditional Random Forests did not achieve good prediction when the correlation between all the variables was low (0.2) in the 100-covariate simulation study: this was a common result to all methods. On the contrary, when the correlation between variables was high (0.8), prediction accuracy was good as expected (similar to MissForest-Lasso), especially in the missing data scenario when Conditional Random Forests followed MissForest imputation (hypothesis 4). Regarding the variable importance performance Random Forests tended to give more importance to continuous variables than binary ones even when they were FP, as expected. This was due to the bias of the Random Forests variable importance measure (Strobl, AL Boulesteix, and Zeileis 2007). Conditional Random Forests did not have this problem and always gave more importance to the TP, as expected (Strobl, Boulesteix, et al. 2008).

- Conditional Random Forests and MissForest imputation accuracy performances improved with increasing correlation between the variables, as expected (hypothesis 4, Tang and Ishwaran 2017). This result extended also to MICE-Lasso and MICE-Elasticnet in prediction accuracy.

- MICE combined with Lasso as for Musoro et al. (2014) did not return good variable selection for the 3% tolerance model. There was an error in the code used by the authors (see Section B.1 in the Appendix) for which their applied tolerance penalty percentages were in truth higher than what they thought for both prediction accuracy and variable selection. Their published 3% tolerance variable selection results were equivalent to my simulation results of the 15% tolerance models. Therefore, their positive results for variable selection and their recommendations are not valid.

- MissForest combined with Lasso did not select the noise terms with large missingness, as expected for Random Forests imputation methods (hypothesis 6, Lu and Petkova

2014). Lasso seemed to counterbalance the known bias of any Random Forest imputation method (Cutler et al. 2009), preventing the noise terms with large missingness (variables $\boldsymbol{X}_3$ and $\boldsymbol{X}_{13}$, see Subsection 2.2.1) from appearing important when the correlation matrix was sparse with high and low correlations. When variables were all strongly or all weakly correlated between each other, this phenomenon diminished.

- MissForest combined with Random Forests tended to give more often importance to the noise variables with large probability of missingness, as expected (hypothesis 7, Cutler et al. 2009). Again this was due to the bias of Random Forest imputation which applies when there is a difference in the prevalence of missing values and in the scale among predictors (Lu and Petkova 2014). When MissForest was combined with Conditional Random Forests, the above problem did not occur and all variables were treated alike, giving a variable selection result better than the other methods.

- MissForest outperformed MICE, when combined with Lasso, in all missing data scenarios, also with increasing ratio FP to TP in the imputation models (i.e. the scenarios with assumption of moderation), as expected (hypothesis 8, Hardt, Herke, and Leonhart 2012). For both MissForest and MICE, the inclusion of many noise variables in the imputation model introduced uncertainty. This negatively affected the prediction and selection performance of MissForest-Lasso in the 100-covariate case where the ratio FP to TP was the highest (234:15). The 20-covariate dataset study showed that having missing data in the outcome and using a more parsimonious imputation model (even excluding some of the TPs, i.e. four interaction terms) gave better MICE-Lasso performance than having a complete outcome with many noisy interaction terms in the imputation model. However, MissForest dealt better with noise than MICE. Moreover, after both Random Forest imputation and MICE, variables that had more missing observations had increased correlation with the outcome variable more than those with less missingness. However, this increment had limited effect on the variable selection of MissForest-Lasso compared to MICE-Lasso.

- Prediction accuracy decreased and optimism estimates increased when missing data were present also in the dependent variable, as expected (hypothesis 9). MissForest-Lasso and MissForest-Random Forests showed better validated prediction when missing data were introduced in the outcome compared to the scenario with complete outcome. MissForest algorithm minimises the error in the RF predictions iteratively until the error cannot be returned smaller (see MissForest algorithm in Subsection 2.1.2). However, the

good validated accuracy of MissForest was too optimistic and did not repeat in the simulated external validation. All other methods accuracy decreased with missing data in the outcome. Moreover, missingness in the outcome caused worse variable selection performance for all methods in the 20-covariate datasets with MissForest-Lasso performing best among the studied methods (I assume that MissForest-Conditional Random Forests would have performed even better as for the 100-covariate scenario results). As the 100-covariate study only explored the scenario with missing data in the outcome, I assume that there would be better prediction accuracy and variable selection when the outcome is complete, similarly to what happened in the 20-covariate case.

- The performances of the studied methods were not equivalent with MAR or MCAR data, contrarily to what I expected (hypothesis 10). The MAR scenarios gave slightly better results in prediction accuracy than the MCAR scenarios, because MAR missingness depends on the other observed variables and the imputation algorithm can account for it (Moritz et al. 2015). This was not convenient for MCAR data in the simulations. MAR data performance was slightly worse than MCAR only in the case of strong correlations between variables and the opposite was true for the low correlation scenario.

Given the results, MissForest-Lasso using the 1 SE and 3% tolerance penalties and MissForest-Conditional Random Forests were the best performing methods in prediction accuracy and MissForest-Conditional Random Forests was better in variable selection (for Random Forests the latter stands for variable importance) in the missing data scenarios.

MissForest-Lasso 1 SE and 3% tolerance models in particular had good results when the average percentage of missing data was 15% (20% or 50%, 12 complete variables), the outcome was complete and there was no assumption of moderation in the 20-covariate dataset scenarios.

In the extreme case of the 100-covariate dataset, in which the ratio noise to active predictors was 234:15 for the moderation assumption, the missing data percentage was approximately 40% (20% or 50%, only 6 complete variables) and there was 20% missingness in the outcome, MissForest-Lasso had poor general performance according to my subjective criteria, even though it had the best prediction accuracy altogether with MissForest-Conditional Random Forests and the second best variable selection result after MissForest-Conditional Random Forest. In this 100-covariate scenario, the 3% tolerance MissForest-Lasso model was preferred to the 1 SE tolerance model for its better variable selection as happened with Lasso in the complete data case.

MissForest-Lasso selected predictor variables well in the low correlation setting but performed poorly in the extreme high correlation setting, even though the selection of only moderators was better than the overall variable selection when covariates were highly correlated. This was expected, as in general Random Forests methods deal with interactions and non-linearities in the data better. Instead, MissForest-Conditional Random Forests was advantageous by better identifying predictor variables using importance measures in both correlation settings (but retaining all variables in the model). However, for clinical practicability MissForest-Lasso would be more useful as the final model will only contain a limited number of predictors chosen from an initial larger number of variables, by reducing the clinicians data collection on new samples. A working prediction model is only of usefulness in clinical practice when the number of variables is small enough to assess a patient in a reasonable time. Selecting a false predictor, which still provides good prediction accuracy due its strong correlation with the true predictor, may therefore be preferable. On the contrary, MissForest-Conditional Random Forests will need all the variables used to develop the model measured by the clinicians in order to make predictions on new patients. Also, it is impossible to express the model linear predictor with an equation as it can be done for regularised regression.

For the purpose to develop a precision medicine model for the psychological treatment CRT with a relative large number of variables and with low to moderate correlations between variables, I prefer MissForest-Lasso 3% tolerance model as the best compromise between prediction accuracy, interpretability and usefulness as a tool in clinical practice. However, if the number of variables relative to sample size increases and/or strong correlation between variables is expected, MissForest-Conditional Random Forests should be considered, e.g. in the analyses of MRI brain imaging data.

### 2.5.1 Advantages and limitations

For the first time, variable selection and bootstrap-validated prediction accuracy for Lasso and Elasticnet have been analysed for different levels of penalty tolerance in a simulation study with different missing data scenarios (MAR and MCAR), missing data percentages (15%, 40%), sample sizes (250, 500 and 1000), number of covariates in the model (20, 40, 249), correlation between covariates (mixed and sparse, all 0.2 and all 0.8), ratios noise to active predictors (10:10, 24:15 and 234:15), with or without missing data in the outcome (20%) and with or without a moderation assumption. All simulations demonstrated that MissForest imputation combined with Lasso produced better results for prognostic and personalized medicine prediction models than MICE, as it better adapts to complex data with non-linearities and interactions.

It also demonstrated that MissForest-Lasso is in a variety of settings is a good compromise between prediction accuracy and interpretability and therefore a first choice for clinicians to develop prediction models. Model performance was not always good according to my criteria, but these refer to final models assessed in clinical practice. In earlier stages, lower model performances are acceptable.

However, I did not run all the scenario combinations and new scenarios, such as intermediate correlations of variables, for all methods for time reasons as the algorithms were computationally expensive in time. Also, having two separate studies (20 and 100 covariate data), with different settings and methods, removes continuity to the interpretation of results that might be true for the analysed scenario-combination and false for the scenario-combination that was not explored. For example, I decided to only run the scenario S5, i.e. with assumption of moderation and missing data also in the outcome for the 100-covariate data simulations. Moreover, in this 100-covariate scenario there were the highest percentage of missing data, the highest ratio FP to TP and the most extreme correlation settings. This mimics our clinical data set best. However, because this simulation scenario S5 performed the poorest in the 20-covariate study for all methods, I assume that the methods in the 100-covariate case would perform better when the outcome is complete.

Furthermore, it would have been better to simulate more complex missing data patterns and correlation scenarios in the large covariate data set, to make the simulations more realistic, especially to decide between MissForest-Lasso and MissForest-Conditional Random Forests models and their variable selection abilities. However, my results suggests that MissForest-Lasso performs well if variables are not highly correlated. Nevertheless, the critical cut-off point is not known for larger number of variables.

# Chapter 3

# Development of MissForest-Lasso prediction model using CRT randomised controlled clinical trial data

## 3.1 Introduction

In the previous chapter 2, I compared the accuracy and variable selection performances of the combined methods MICE-Lasso, MICE-Elasticnet, MissForest-Lasso, MissForest-Elasticnet, MissForest-Random Forests and MissForest-Conditional Random Forests through simulations. The method showing the best trade-off between accuracy and variable selection performance was MissForest-Lasso.

In this chapter, I will apply MissForest-Lasso to identify moderators of Cognitive Remediation Therapy (CRT) heterogeneity in patients with schizophrenia in order to develop prediction models of treatment success. I will use individual participant data from multiple randomized controlled trials (RCT) assessing the effectiveness of CRT. These precision medicine models will use characteristics of patients measured before treatment (baseline variables) and interactions between baseline variables and treatment type as predictors.

In this introductory section, I will write about the Database of Cognitive Training and Remediation Studies (DoCTRS) and the CIRcuiTS Combined Data (CCD) used to build the model. Then I will describe how I merged the databases and present the pooled database summary statistics and missing data.

The second section of the chapter will explain the methodology for the development of the prediction models. I will develop three models, one using a summary score of cognitive ability measures of memory, processing speed and executive function as treatment outcome and two using the Wisconsin Card Sorting Test Perseverative Errors (WCST PE) measure of executive function as treatment outcome. The summary measure will be computed through factor analysis in order to predict altogether different end-of-treatment cognitive abilities with an univariate prediction model (i.e. allowing for single dependent variable). WCST PE, one of the clinical most relevant measures, was chosen as the dependent variable of the other models in order to apply the model to an observed outcome instead of a latent measure. The two models with WCST PE as the dependent variable will be run as follows: one model will only use the complete outcome cases, the other model will also include the patients with missing outcome. There will be no missing outcome data for the latent variable model. Finally, to assess the predictive power and influence of identified moderators, I will also develop prediction models without the assumption of moderation of treatment outcome.

The third section will be dedicated to the results and comparison of the predictive performances of the developed models.

### 3.1.1 DoCTRS randomised controlled trials

Individual participant data from nine different RCTs were used to develop and validate the prediction model. The datasets available were from the Database of Cognitive Training and Remediation Studies (DoCTRS, containing only seven studies at the time when the PhD project began), the Computerised Interactive Remediation of Cognition Training for Schizophrenia or CIRcuiTS Combined Data (CCD, one study) and two later DoCTRS RCTs by Fiszdon et al. (2016):

1. The DoCTRS is run by NIMH in the USA and it is an attempt to open source data/data sharing. At the time of the start of the PhD, the DoCTRS comprised data from seven different RCTs called after their investigators: ‘Wykes 1’ (Wykes, Reeder, Corner, et al. 1999), ‘Bell’, (Bell et al. 2008), ‘Keefe’ (Keefe et al. 2012), ‘Wykes 2’ (Wykes, Reeder, Landau, Everitt, et al. 2007), ‘Wykes 3’ (Wykes, Newton, et al. 2007), ‘Keshavan’ (Eack et al. 2009) and ’Silverstein’ (Silverstein et al. 2009). To be eligible for all of the studies, a patient needed to be diagnosed with SCZ according to the Diagnostic and Statistical Manual of Mental Disorders (DSM) and the International Classification of Diseases (ICD) criteria. In three studies (‘Bell’, ‘Keefe’ and ‘Keshavan’) the eligibility criteria extended to

people with schizoaffective disorder. Two studies ('Keefe' and 'Keshavan') required the patients to be aged between 18 and 55; one study ('Wykes 2') accepted people with a minimum age of 17 and the remaining did not have age eligibility criteria.

The DoCTRS contained five datasets of variables measured on 430 patients and one dataset regarding the study information (the list of variables can be found in the Appendix C). Each variable in the datasets (apart from the demographics and the study information datasets) was measured up to six different time points per person (screening, baseline, two midpoints, end-of-treatment and follow-up). The datasets were the following:

- Study information (see Table 3.1): 192 variables providing information about heterogeneities and similarities of the different studies, database divided in 13 sections:
  - General study information
  - Summary of subject and study characteristics
  - Intervention and comparison condition characteristics
  - Techniques used in cognitive remediation intervention
  - Treatment targets
  - Cognitive remediation delivery methods
  - Cognitive remediation delivery format
  - Cognitive remediation delivery setting
  - Other treatment features
  - Defining "completion"
  - Eligibility criteria
  - Summary of study assessment schedule
- Cognitive data: 38 cognitive outcomes (12 memory outcomes, one processing speed outcome and 25 executive function outcomes, see Table 3.2), 9 global cognition variables and 9 other types variables (not to be used as dependent variables but included as covariates in the prediction model) and six variables without any record (null variables). The six occasion measurements for the variables were not available for all studies. All non-null variables were available as raw and scaled versions apart from four outcome variables that were only available in a scaled version (t-score).
- Demographics: general patients baseline characteristics (36 variables) such as age, gender,ethnicity, racial category, marital status, education years, education category, primary psychiatric diagnosis, age of onset of psychiatric symptoms and age of first treatment for psychiatric symptoms;

- Medications (MED): type and dose of antipsychotics used (30 variables with five time points' measures per patient – only one midpoint)
- Functioning, self-esteem and quality of life measures: measures of functional outcomes and wellbeing (59 variables with five time points' measurements per person). Again not all the studies had the five time measurements available. The measures were: Rosenberg Self-esteem Scale (RSE, alternative scoring, three studies), Heinrich-Carpenter Quality of Life Scale (HCQOL, two studies), Social Adjustment Scale II (SAS-II, only 'Keshavan') and Social Behaviour Scale (SBS, three studies).
- Symptom data: positive and negative symptom measures (131 variables measured on five different occasions not available for all studies). The measures were: Positive and Negative Syndrome Scale (PANSS, Kay, Fiszbein, and Opler 1987, two studies), Brief Psychiatric Rating Scale (BPRS, Overall and Gorham 1962, two studies), Scale for the Assessment of Negative Symptoms (SANS, Andreasen 1982, only 'Bell') and Scale for the Assessment of Positive Symptoms (SAPS, Andreasen 1984, only 'Bell').

2. The CCD (Cella, Bishara, et al. 2014, Reeder et al. 2017) was collected for a RCT from two centres: CIR01 London and CIR02 Sussex for a total of 120 patients and 30 baseline variables and 310 variables measured at three time points: baseline, end-of-treatment and follow-up. Eligibility criteria were: DSM diagnosis of SCZ or schizoaffective disorders, no less than one year's contact with mental health services, 17 to 65 years old and performance more than one SD below the normative mean in working memory. Only ten of the variables were cognitive outcomes (three memory outcomes, one processing speed and six executive function outcomes, see Table 3.2). About 280 variables were common to DoCTRS, of which eight were common outcomes (one memory, one processing speed and six executive function outcomes, see Table 3.2) and 192 were all the study information variables. The longitudinal variables for this dataset were recorded for all the three time points. Common measures of quality of life and symptom data were SBS and PANSS respectively.

3. The two RCTs investigated by Fiszdon were titled: 'Predictors of response to cognitive remediation in SCZ' (75 individuals, (2016) and 'Efficacy of Social Cognition Training in Schizophrenia' (52 individuals, unpublished). The participants had confirmed diagnosis of schizophrenia spectrum disorder (schizophrenia, schizo-affective disorder, psychosis not otherwise specified or affective disorder with psychotic features) and were aged 18–65

years. About 20 variables were common to DoCTRS in each study.

The first study, 'Fiszdon 1', had eight common outcomes to DoCTRS (two memory measures, i.e the California Verbal Learning Test-II (CVLT) and Wechsler Memory Scale (WMS); one processing speed outcome, i.e. the Trail Making Test Part A (TMTA) and four executive function outcomes, i.e the Wisconsin Card Sorting Test (WCST) measuring % perseverative errors and % conceptual level, the Trail Making Test Part B (TMTB) and the verbal fluency test (FAS), see Table 3.2). All the study information variables were in common. The longitudinal variables for this dataset were recorded for three time points. In the data common measures of symptoms were the five-factor PANSS variables.

The second RCT, 'Fiszdon 2', had only three common outcomes to DoCTRS (one memory outcome, i.e. Letter number span (LNS), one processing speed outcome, i.e. Trail Making Test Part A (TMTA) and one executive function outcome, i.e. Category fluency Animal naming (CATFLU), see Table 3.2). Longitudinal variables were only measured at two time points: baseline and end-of-treatment. However, 23 of this study participants were common to the first study, this implying their exclusion from any analysis using the first study whole data.

DoCTRS and CCD were used for model development and the Fiszdon study data for the model external validation. Fiszdon's data were used as validation set because they became available later during the PhD with other DoCTRS data (the latter had less common variables to the model development data compared to Fiszdon).

**Studies' description** The studies 'Wykes 1,2 and 3' and 'Circuits' were conducted in the UK, while the other were conducted in the USA. All the studies' data were previously analysed singularly reporting some significant ('Wykes 1', Wykes, Reeder, Corner, et al. 1999, 'Wykes 2', Wykes, Reeder, Landau, Everitt, et al. 2007, 'Wykes 3', Wykes, Newton, et al. 2007, 'Keshavan', Eack et al. 2009 and 'Fiszdon', Fiszdon et al. 2016) or borderline significant improvement ('Circuits', Reeder et al. 2017) for some of the single cognitive outcomes after receiving CRT. Namely, the study 'Wykes 1' (Wykes, Reeder, Corner, et al. 1999) and 'Wykes 3' (Wykes, Newton, et al. 2007) showed that executive function significantly improved after CRT and the outcome used was the number of WCST categories achieved (see Table 3.2).

The RCT 'Wykes 2' (Wykes, Reeder, Landau, Everitt, et al. 2007) presented a significant result for the outcome Wechsler Adult Intelligence Scale (WAIS) working memory digit span test (see Table 3.2), indicating that memory was enhanced after treatment. Keefe et al. (2012)

used the MATRICS Consensus Cognitive Battery (MCCB, see Table 3.2) to measure cognitive function in their study and there was a significant effect of CRT only at midpoint, but not at the end-of-treatment. The memory outcome California Verbal Learning Test (CVLT) short-term free recall and the executive function outcomes TMTB and Tower of London (TOL) Ratio of initiation to Execution time (see Table 3.2) were tested in RCT 'Keshavan' (Eack et al. 2009) and significant improvements were found for each outcome after CRT. Finally, in the study 'Circuits' (Reeder et al. 2017) the tested executive function outcome WCST perseverative errors revealed a borderline significant effect of CRT.

**Merging process and data preprocessing**

I merged the source databases DoCTRS and CCD into one single target database for the model building process by using all the available variables (variables in the final database could also be only measured for one study and one time point). The DoCTRS 'Silverstein' study was excluded from our analysis because it did not have any measure of cognitive outcomes, meaning that 82 patients out of 550 were dropped from the pooled database.

In the next step, I familiarized myself with the structures of target and validation datasets to clean the data sets and to harmonize the available variables. Data cleaning and inspection revealed a variety of issues which had to be solved by consulting the clinical expert in my project:

- 19 patients were taking antipsychotics with more than 5000 mg of chlorpromazine equivalent per day, which is a lethal dose and therefore not plausible. Thus, in agreement with the clinician, either I considered these doses as missing or substituted them with the highest likely value under 5000 mg of chlorpromazine equivalent present in the data (2833 mg) depending on the plausibility of the other records for the patients having these high doses.

- Variables in DoCTRS had more or different categories compared to the corresponding variables in CCD. Therefore, I harmonized the different levels of these categorical variables (for example 'race', 'marital status', 'education level achieved') by collapsing several levels into one. Also, some common variables were continuous in one database and categorical in the second: these were merged as categorical or ordinal variables (e.g. time since/age of first treatment for psychiatric symptoms, time since/age of first psychiatric hospitalization).

- I created several variables that were already present in CCD but not in some studies of DoCTRS by combining two other variables in DoCTRS (for example: the cognitive outcome Wisconsin Card Sorting Test (WCST) percentage of total error already present in CCD was created in DoCTRS by combining WCST non-perseverative errors with WCST perseverative errors; the chlorpromazine equivalent per day values were computed from the daily doses of each antipsychotic taken by patients, etc).

Therefore, the final merged corrected database for model training had data from seven studies (see summary statistics of most important variables in Table 3.3): 'Wykes 1, 2 and 3', 'Bell', 'Keefe', 'Keshavan' and 'Circuits'. The overall sample size was 468 and the number of variables was 547. The variables were measured in up to six different occasions, thus a wide format of the database contained 2896 time-point specific variables, of which 1706 were variables with at least one record, and 2808 were the person-by-time observations.

The cognitive outcome variables were in total 38: 12 describing 'memory', one processing speed and 25 'executive function', only four of them were common to the two databases, three of which were common to six studies (i.e. Wechsler Adult Intelligence Scale (WAIS) Digit Span, memory measure; WCST percentage of total error and WCST perseverative errors, executive function measures); and the last one was common to only five studies: WCST categories achieved).

### Summary baseline statistics

Summary characteristics for important demographics, symptoms, quality of life, global cognition baseline putative predictors and baseline cognitive outcomes per study are provided in Table 3.3 (only the statistics for the 7 studies data used to train the model are shown).

**Study characteristics** Study sample sizes were small (mean=68, range=40–120). Four studies were carried out in the United Kingdom (the 3 'Wykes' and 'Circuits') and the others were done in the United States. The average number of cognitive domains measured per study was 14.1, with a range of 10–20. The studies 'Wykes 3' and 'Keshavan' had on average the youngest patients (see Table 3.3). All studies had a larger proportion of male participant than female (mean percentage across studies 69%, SD 7) and on average 12.44 participant years of education per study (SD 0.91, 6 studies).

**Overall sample characteristics** The study samples used for model training consisted of individuals with age ranging from 14 to 66 (overall mean age = 34.81 years [SD = 11.50], see

Table 3.3 for study-specific summary statistics) who were mostly men (68%) with 12.60 years of education (range = 2–20 in six studies). Positive, negative and general symptom severity (as measured by PANSS summary scores), when reported (three studies), was mainly in the minimal to moderate range. 'Wykes 1' and 'Wykes 3's measures of symptoms through BPRS were also mild. Within-study ranges suggested that some studies included more severely symptomatic individuals.

The 75 participants of the study 'Fiszdon 1' were used for model external validation. Individuals were on average in their late forties (47.8, range 27-64) with 12.4 years of education (range 7-16). Symptoms severity was minimal to mild according to PANSS measures (mean PANSS total = 52.7, range 31-93).

**Treatment characteristics** All studies compared cognitive remediation therapy (CRT, see Subsection 1.1.1) to treatment as usual (TAU). As reported by some of the studies, TAU consisted of routine psychiatric care delivered within the UK National Health Service. Community, inpatient or rehabilitation wards were the settings for TAU. TAU included medication review and monitoring by psychiatrist, meetings with a mental health nurse for support, attendance at day centres, some occupational therapy, computer games, residential support with self-care or rehabilitation programmes (Keefe et al. 2012, Eack et al. 2009 and Reeder et al. 2017)

There were some differences in how CRT and TAU were delivered between studies (see Table 3.1). For example, only the study 'Bell' used additional employment and vocational rehabilitation in both CRT and TAU. Moreover 'Bell' also used the intervention life style group in the CR condition. The study 'Keefe' did not offer psycho-medicine management in CRT, contrarily to the other studies. The drill and practice technique was central to the intervention in the studies: 'Bell', 'Keefe' and 'Circuits'. The UK studies gave more importance to the metacognitive training and errorless learning techniques than the US studies. In vivo practice techniques were central to the intervention in 'Keshavan'. Four studies involved the use of drill and practice exercises on a computer ('Bell', 'Keefe', 'Keshavan' and 'Circuits'). CR delivery format was one-on-one for all studies but 'Keefe' and 'Keshavan', which had a mix of individual and group sessions. The typical duration of a CR session was 60 minutes (90 minutes for 'Keshavan') delivered on average 4.8 times per week (range 2-10) for a mean duration of CR intervention of 26.5 weeks (range 8-96, 12 for the UK studies). CR trainers were all research staff (doctoral, master or non-graduated level) apart from 'Circuits's staff that were master-level clinicians. Only three studies, 'Bell', 'Keefe' and 'Fiszdon 1' paid participants to undergo both therapy and assessment sessions. In 'Keshavan' and 'Circuits', participants were only paid for

assessments.

### Missing data

There was a substantial amount of missing data in the pooled database: missingness by design (some studies did not have screening and mid-points measures by design, data are MCAR) and missingness due to drop-outs and intermittent missing values (e.g. some patients had the follow-up measure, but not the end-of-treatment measure). The overall mean percentages of missing data across studies per time point (including missingness by design) were the following: screening 95%, baseline 73%, 1st mid-point 92%, 2nd mid-point 96%, end-of-treatment 75%, and follow-up 78%. Missingness per time points across the studies where variables were measured (i.e. excluding missingness by design) was: screening 39% (SD 44, study 'Keefe'), baseline 11% (SD 26, all studies), 1st mid-point 24% (SD 15, studies 'Keefe', 'Keshavan' and 'Bell'), 2nd mid-point 84% (SD 30, study 'Bell'), end-of-treatment 18% (SD 23, all studies) and follow-up 26% (SD 25, all studies). The missing-data patterns are similar between treatment groups and they are heterogeneous between studies. There were overall 76% completers and a completer was defined as a patient having had at least 20 hours of therapy received.

| VARIABLES | STUDIES | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **B** | **C** | **Kf** | **Ks** | **W1** | **W2** | **W3** | **F1** |
| Study country | USA | UK | USA | USA | UK | UK | UK | USA |
| Other CR approach used in CRT? | Yes | No | No | No | No | No | No | No |
| Other intervention used in comparison condition? | No | No | No | Yes | Yes | Yes | Yes | No |
| Patients in CR condition get antipsychotics management? | Yes | Yes | No | Yes | Yes | Yes | Yes | Yes |
| Rank order of Drill and Practice technique in CRT | 1 | 3 | 1 | NC | NC | NC | NC | 1 |
| Rank order of Strategy training technique in CRT | 1 | 2 | 2 | 4 | 2 | 3 | 3 | 2 |
| Rank order of metacognitive training technique in CRT | NC | 1 | NC | NC | 2 | 1 | 1 | NC |
| Rank order of errorless learning technique in CRT | 2 | 4 | NC | NC | 3 | 3 | 3 | NC |
| Rank order of general cognition among targets of CRT | 1 | 1 | 1 | 2 | 3 | 2 | 2 | 1 |
| Rank order of attention among targets of CRT | 3 | NP | NP | NP | 4 | 4 | 4 | 2 |
| Rank order of verbal memory among targets of CRT | 4 | 2 | NP | NP | 2 | 3 | 3 | 3 |
| Rank order of other target among targets of CRT | NP | NP | NP | 3 | 1 | 1 | 1 | NP |
| CR sessions delivered one-on-one? | Yes | Yes | No | No | Yes | Yes | Yes | Yes |
| Typical duration of CR session (minutes) | 60 | 60 | 60 | 90 | 60 | 60 | 60 | 60 |
| Target number of CR sessions per week | 10 | 3 | 5 | 2 | 4 | 5 | 4 | 5 |
| Duration of CR intervention (in weeks) | 52 | 12 | 8 | 96 | 12 | 12 | 12 | 8 |
| Duration and frequency of control condition similar to CR? | No | Yes | Yes | No | Yes | No | No | Yes |
| Did doctoral-level clinicians administer CR? | Yes | No | No | Yes | No | No | No | NA |
| Did trainers without graduate training administer CR? | Yes | No | Yes | No | Yes | Yes | Yes | NA |
| Were CR trainers research staff? | Yes | No | Yes | Yes | Yes | Yes | Yes | NA |
| Patients with diagnosis of schizoaffective disorder eligible? | Yes | Yes | Yes | Yes | No | No | No | Yes |
| Weeks between baseline and end-of-treatment assessment | 52 | 12 | 12 | 104 | 12 | 12 | 12 | 8 |
| Weeks between post-treatment and 1st follow-up assessment | 52 | 12 | NA | NA | 24 | 24 | 12 | 8 |

Table 3.1: Study information variables: the most important. Abbreviations: B='Bell', C='Circuits', Kf='Keefe', Ks='Keshavan', W1='Wykes 1', W2='Wykes 2', W3='Wykes 3', F1='Fiszdon 1', CRT=cognitive remediation therapy, NC=non central to the intervention, NP=no priority target, NA=missing

Table 3.2: **Outcomes variables** details. Abbreviations as follows: R=raw score, T=t-score, M=memory, EF=executive function, PS=processing speed, Cont=continuous, Cat=categorical, W1='Wykes 1', B='Bell', Kf='Keefe', W2='Wykes 2', W3='Wykes 3', Ks='Keshavan', C='Circuits', MD=missing data, b=baseline, e=end-of-treatment, f=follow-up.

| Outcome | Description | Measure | Scale | Studies | Overall MD(%) | | |
|---|---|---|---|---|---|---|---|
| | | | | | b | e | f |
| BVMTR | Brief Visuospatial Memory Test - Revised (included in the MCCB) 3-trial total recall; R. | M | Cont | Kf | 90 | 89 | 100 |
| CATFLU | Category fluency: Animal naming (n animals named in 60 s); R | EF | Cont | B, Kf, W3 | 65 | 69 | 84 |
| CPT IP | Continuous Performance Test - Identical Pairs (Mean across 2-, 3-, and 4-digit conditions; R. | EF | Cont | Kf | 90 | 89 | 100 |
| CVLT LFR | California Verbal Learning Test: Long-term Free Recall | M | Cont | Ks | 88 | 90 | 100 |
| CVLT SFR | California Verbal Learning Test: Short-term Free Recall | M | Cont | Ks | 88 | 90 | 100 |
| CVLT TR | California Verbal Learning Test: Total Recall | M | Cont | Ks | 88 | 90 | 100 |
| FAS A | Verbal fluency (FAS): Age- and education-adjusted score | EF | Cont | W1, W2, W3 | 66 | 69 | 72 |
| FAS NR | Verbal fluency (FAS): Total number of correct responses | EF | Cont | W1, W2, W3, B | 50 | 56 | 63 |
| HAY A | Hayling section 2 category A errors (A score) | EF | Cont | W2, W3, C | 75 | 76 | 77 |
| HAY B | Hayling section 2 category B errors (B score) | EF | Cont | W2, W3, C | 75 | 76 | 77 |
| HAY T | Hayling total scaled score (section 1 + section 2 + section 2 errors scaled scores) | EF | Cat | W2, C | 49 | 70 | 70 |
| HAY TE | Hayling overall scaled score | EF | Cont | W2, W3, C | 66 | 53 | 56 |
| LNS | Letter-Number Span (n of correct trials); R. | M | Cont | Kf, W3, Ks | 63 | 66 | 78 |
| MCCB Ve | MATRICS Verbal Learning domain; T. | M | Cont | Kf | 90 | 89 | 100 |
| MCCB Vi | MATRICS Visual Learning domain; T. | M | Cont | Kf | 90 | 89 | 100 |
| MCCB WM | MATRICS Working Memory domain; T. | M | Cont | Kf | 90 | 89 | 100 |
| MCCB RP | MATRICS Reasoning and Problem Solving domain; T. | EF | Cont | Kf | 90 | 89 | 100 |
| MSET A | Modified six elements task: Number of tasks attempted | EF | Cat | W1, W2, W3 | 66 | 70 | 72 |
| MSET R | Modified six elements task: Number of rules broken | EF | Cat | W1, W2, W3 | 66 | 70 | 72 |
| MSET T | Modified six elements task: Total score: no. of tasks attempted - no. of rule breaks | EF | Cat | W1, W2, W3 | 66 | 70 | 72 |
| NAB M | Neuropsychological Assessment Battery: Mazes subtest; R. | EF | Cont | Kf | 90 | 89 | 100 |
| REY T2 | Rey complex figure test immediate recall; R. | M | Cont | C | 74 | 76 | 76 |
| REY T3 | Rey complex figure test delayed recall; R. | M | Cont | C | 74 | 76 | 76 |
| TMTA | Trailmaking test part A (Paper & pencil): time to completion (seconds) | PS | Cont | B, Kf, W2, W3 | 47 | 52 | 69 |
| TMTB | Trailmaking test Part B (Paper & pencil): time to completion (seconds) | EF | Cont | B,W2, W3, Ks | 46 | 54 | 69 |
| TMTB E | Trailmaking test Part B (Paper & pencil): Number of errors | EF | Cat | B, W2, W3 | 58 | 70 | 69 |
| TMTB C | Trailmaking test Condition 2/letters+numbers (Computerized): trial 1, time to completion | EF | Cont | W1 | 92 | 93 | 94 |
| TOLDX M | Tower of London - DX: Total move score (range: 0 to 189) | EF | Cont | Ks | 87 | 90 | 100 |
| TOLDX IE | Tower of London - DX: Ratio of initiation to Execution time (range: 0 to 1) | EF | Cont | Ks | 87 | 90 | 100 |
| WAIS DG | Wechsler Adult Intelligence Scale Digit Span; R. | M | Cont | W1, B, W2, W3, Ks, C | 12 | 23 | 39 |
| WAIS PA | Wechsler Adult Intelligence Scale Picture Arrangement; R. | EF | Cont | W1, B, Ks | 64 | 70 | 84 |
| WAIS PC | Wechsler Adult Intelligence Scale Picture Completion; R. | EF | Cont | W1, B, W3 | 68 | 86 | 87 |
| WCST C | Wisconsin Card Sorting Test: Categories Achieved (0 to 6) | EF | Cat | W1, B, W2, W3, C | 25 | 33 | 40 |
| WCST NE | Wisconsin Card Sorting Test: Non-Perseverative Errors (0 to 128) | EF | Cont | W1, B, W2, W3, Ks | 37 | 47 | 63 |
| WCST PC | Wisconsin Card Sorting Test: Percent Conceptual Responses (0 to 100) | EF | Cont | W1, B, W2, W3, Ks | 37 | 47 | 63 |
| WCST PE | Wisconsin Card Sorting Test: Perseverative Errors (0 to 128) | EF | Cont | W1, B, W2, W3, Ks, C | 13 | 24 | 40 |
| WCST TE | Wisconsin Card Sorting Test: Percentage Total Errors; R. | EF | Cont | W1, B, W2, W3, Ks, C | 13 | 25 | 40 |
| WMS | Wechsler Memory Scale; III: spatial span subtest (sum of scores on backwards and forwards conditions); R. | M | Cont | Kf | 90 | 89 | 100 |

Table 3.3: **Summary baseline characteristics**: 24 baseline variables with up to 66% missing data across studies are here summarised (the first 18 variables are some of the predictors and the remaining 6 are some of the outcomes). Means (SD) or counts (%) with an asterisk are computed in presence of missing data (i.e. by excluding the missing values). Abbreviations: CRT=cognitive remediation theraphy, Demo=demographics, M=mean, SCZ=schizophrenia, T=treatment, H=hospitalization, QoL=quality of life, IQ=intelligent quotient, P=perseverative; for PANSS, SBS, WAIS, WCST, TMTA/B, LNS, FAS and CATFLU see Table 3.2.

| **VARIABLES** | | **STUDIES** | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Mean (SD), Count (%) | **Bell** | | | **Circuits** | | | **Keefe** | | | **Keshavan** | | | **Wykes 1** | | | **Wykes 2** | | | **Wykes 3** | | |
| **Predictors** | | CRT | Control | All | CRT | Control | All | CRT | Control | All | CRT | Control | All | CRT | Control | All | CRT | Control | All | CRT | Control | All |
| Demo. | Sample size | 42 | 35 | 77 | 59 | 61 | 120 | 27 | 26 | 53 | 31 | 27 | 58 | 18 | 17 | 35 | 43 | 42 | 85 | 21 | 19 | 40 |
| | Age | 41.12 (9.98) | 37.37 (8.93) | 39.42 (9.64) | 39.98 (11.24) | 37.64 (10.93) | 38.79 (11.10) | 36.07 (10.27) | 37.92 (10.38) | 36.98 (10.27) | 25.42 (6.50) | 25.67 (6.33) | 25.53 (6.37) | 36.56 (7.57) | 41.06 (9.97) | 38.74 (8.98) | 36.67 (11.61) | 35.69 (9.03) | 36.19 (10.37) | 18.76 (2.57) | 17.47 (2.17) | 18.15 (2.44) |
| | Male | 27 (64.29) | 17 (48.57) | 44 (57.14) | 43 (72.88) | 37 (60.66) | 80 (66.67) | 22 (81.48) | 17 (65.38) | 39 (73.58) | 20 (64.52) | 20 (74.07) | 40 (68.97) | 14 (77.78) | 13 (76.47) | 27 (77.14) | 26 (60.47) | 36 (85.71) | 62 (72.94) | 13 (61.90) | 13 (68.42) | 26 (65.00) |
| | Not white | 24 (57.14) | 17 (48.57) | 41 (53.25) | 34 (57.63) | 41 (67.21) | 75 (62.50) | 8* (29.63)* | 11* (42.31)* | 19* (35.85)* | 10 (32.26) | 8 (29.63) | 18 (31.03) | 2 (11.11) | 5 (29.41) | seven (20.00) | 11 (25.58) | 12 (28.57) | 23 (27.06) | 9 (42.86) | 11 (57.89) | 20 (50.00) |
| | Not married | 38 (90.48) | 33 (94.29) | 71 (92.21) | 55 (93.22) | 55 (90.16) | 110 (91.67) | 23 (85.19) | 26 (100) | 49 (92.45) | NA NA | NA NA | NA NA | 17 (94.44) | 17 (100) | 34 (97.14) | 39 (90.70) | 41 (97.62) | 80 (94.12) | 21 (100) | 19 (100) | 40 (100) |
| | M. Education years | 12.86 (2.67) | 12.49 (1.76) | 12.69 (2.30) | 13.51 (2.62) | 13.16 (2.36) | 13.33 (2.49) | 13.37 (2.37) | 13.62 (2.14) | 13.49 (2.24) | NA NA | NA NA | NA NA | 12.14 (1.94) | 12.59 (2.09) | 12.36 (2.00) | 11.88 (1.75) | 11.06 (2.42) | 11.48 (2.14) | 11.50 (1.90) | 11.08 (1.27) | 11.30 (1.62) |
| | 1st diagnosis SCZ | 29 (69.05) | 20 (57.14) | 49 (63.64) | 47 (79.66) | 47 (77.05) | 94 (78.33) | 18 (66.67) | 21 (80.77) | 39 (73.58) | 21 (67.74) | 17 (62.96) | 38 (65.52) | 18 (100) | 17 (100) | 35 (100) | 43 (100) | 42 (100) | 85 (100) | 21 (100) | 19 (100) | 40 (100) |
| | >10 years since 1st T. | NA NA | NA NA | NA NA | 43 (72.88) | 31 (50.82) | 74 (61.67) | 14 (51.85) | 20 (76.92) | 34 (64.15) | NA NA | NA NA | NA NA | NA NA | NA NA | NA NA | 22 (51.16) | 22 (52.38) | 44 (51.76) | NA NA | NA NA | NA NA |
| | >10 yrs since 1st H. | 26* (61.90)* | 25* (71.43)* | 51* (66.23)* | 36 (61.02) | 27 (44.26) | 63 (52.50) | NA NA | NA NA | NA NA | NA NA | NA NA | NA NA | NA NA | NA NA | NA NA | 19* (44.19)* | 16* (38.10)* | 35* (41.18)* | NA NA | NA NA | NA NA |
| Symptoms | M. PANSS positive | 18.05 (5.72) | 19.37 (5.35) | 18.65 (5.56) | 12.71* (5.2)* | 12.25* (4.78)* | 12.48* (4.98)* | NA NA | NA NA | NA NA | NA NA | NA NA | NA NA | NA NA | NA NA | NA NA | 14.19* (5.41)* | 12.5* (4.99)* | 13.37* (5.24)* | NA NA | NA NA | NA NA |
| | M. PANSS negative | 18.48 (6.30) | 19.74 (5.46) | 19.05 (5.93) | 12.98* (6.23)* | 13.18* (5.77)* | 13.09* (5.97)* | NA NA | NA NA | NA NA | NA NA | NA NA | NA NA | NA NA | NA NA | NA NA | 18.62* (7.30)* | 16.62* (7.23)* | 17.65* (7.29)* | NA NA | NA NA | NA NA |
| | M. PANSS general | 39.36 (8.42) | 39.83 (7.51) | 39.57 (7.97) | 30.11* (8.30)* | 29.12* (8.20)* | 29.6* (8.23)* | NA NA | NA NA | NA NA | NA NA | NA NA | NA NA | NA NA | NA NA | NA NA | 30.07* (8.20)* | 27.55* (6.83)* | 28.84* (7.62)* | NA NA | NA NA | NA NA |
| QoL, self-esteem, functioning | Mean SBS total | NA NA | NA NA | NA NA | 11.05* (8.84)* | 10.53* (7.42)* | 10.79* (8.11)* | NA NA | NA NA | NA NA | NA NA | NA NA | NA NA | 13.33 (7.74) | 11.12* (8.07)* | 12.29* (7.85)* | 11.6 (8.45) | 13.17* (11.21)* | 12.37* (9.86)* | 12.76 (9.13) | 14.44* (9.12)* | 13.54* (9.04)* |
| | M. RSE confirmation | NA NA | NA NA | NA NA | NA NA | NA NA | NA NA | NA NA | NA NA | NA NA | NA NA | NA NA | NA NA | 16.83 (4.48) | 16.12* (5.45)* | 16.50* (4.89)* | 17.33 (4.40) | 16.68* (4.20)* | 17.01* (4.29)* | 16 (4.14) | 18.21 (3.77) | 17.05 (4.07) |
| | M. RSE deprecation | NA NA | NA NA | NA NA | NA NA | NA NA | NA NA | NA NA | NA NA | NA NA | NA NA | NA NA | NA NA | 15.28 (5.30) | 13.19* (3.64)* | 14.29* (4.65)* | 16.21 (4.49) | 15.22* (4.66)* | 15.73* (4.57)* | 16.14 (3.66) | 16.58 (4.81) | 16.35 (4.20) |
| Medications | M. Chlorpromazine | 699.20* (372.52)* | 753.17 (615.33) | 724.05* (496.38)* | 426.83* (362.77)* | 538.92* (425.64)* | 482.88* (397.73)* | NA NA | NA NA | NA NA | 381.18 (271.99) | 460.68 (335.05) | 418.19 (302.89) | 833.78 (572.57) | 910.29 (587.86) | 870.94 (572.76) | 292.95 (307.29) | 259.2 (321.63) | 276.28 (313.04) | 260.89 (224.84) | 253.38 (206.87) | 257.32 (213.75) |
| Global cognition | Mean IQ | NA NA | NA NA | NA NA | 88.41 (12.72) | 86.56 (15.03) | 87.47 (13.91) | NA NA | NA NA | NA NA | 97.74 (7.66) | 98.52 (9.74) | 98.1 (8.62) | NA NA | NA NA | NA NA | NA NA | NA NA | NA NA | NA NA | NA NA | NA NA |
| | M. WAIS vocabulary | 38.00 (15.08) | 34.29 (16.44) | 36.31 (15.72) | 35.08 (13.53) | 31.77 (14.77) | 33.40 (14.21) | NA NA | NA NA | NA NA | 9.97 (2.66) | 9.56 (2.82) | 9.78 (2.72) | 36.35* (14.89)* | 38.75* (18.96)* | 37.52* (16.76)* | 26.98 (13.98) | 28.05 (14.00) | 27.51 (13.92) | 30.52 (12.37) | 30.11 (12.57) | 30.32 (12.30) |
| | M. WAIS digit-symbol | 51.00 (14.02) | 50.14 (17.11) | 50.61 (15.40) | 44.93 (15.17) | 43.25 (16.96) | 44.08 (16.06) | NA NA | NA NA | NA NA | 8.61 (2.47) | 8.00 (2.60) | 8.33 (2.53) | NA NA | NA NA | NA NA | NA NA | NA NA | NA NA | 40.71 (8.74) | 40.78* (12.17)* | 40.74* (10.32)* |
| **Cognitive outcomes** | Mean WCST P. errors | 22.48 (18.22) | 28.17 (21.41) | 25.06 (19.81) | 28.12* (17.30)* | 27.84* (16.57)* | 27.98* (16.87)* | NA NA | NA NA | NA NA | 12.45 (9.72) | 11.15 (7.85) | 11.84 (8.85) | 35.76* (18.44)* | 44.44* (27.42)* | 39.97* (23.28)* | 36.49 (24.70) | 41.62 (26.53) | 39.02 (25.60) | 27.48 (19.63) | 30* (15.88)* | 28.64* (17.81)* |
| | Mean TMTA | 47.02 (22.81) | 44.06 (24.39) | 45.68 (23.43) | NA NA | NA NA | NA NA | 35.05* (12.11)* | 39.54 (20.83) | 37.53* (17.45)* | NA NA | NA NA | NA NA | NA NA | NA NA | NA NA | 62.20 (33.08) | 71.89 (79.60) | 66.99 (60.53) | 46.81 (22.38) | 48.97 (21.12) | 47.83 (21.54) |
| | Mean TMTB | 124.64 (98.01) | 115.73* (59.82)* | 120.72* (83.00)* | NA NA | NA NA | NA NA | NA NA | NA NA | NA NA | 66.29 (26.09) | 69.31* (34.94)* | 67.67* (30.20)* | NA NA | NA NA | NA NA | 174.97* (143.73)* | 164.08* (130.39)* | 169.66* (136.63)* | 136.05 (118.07) | 117.34 (52.56) | 127.16 (92.27) |
| | Mean LNS | NA NA | NA NA | NA NA | NA NA | NA NA | NA NA | 12.48* (3.70)* | 12.31 (4.05) | 12.38* (3.85)* | NA NA | NA NA | NA NA | NA NA | NA NA | NA NA | 6.79* (2.97)* | 6.64 (3.22) | 6.71* (3.08)* | 8.14 (3.02) | 7.74 (2.84) | 7.95 (2.91) |
| | Mean FAS responses | 31.81 (11.58) | 30.46 (10.70) | 31.19 (11.14) | NA NA | NA NA | NA NA | NA NA | NA NA | NA NA | NA NA | NA NA | NA NA | 27.89 (14.49) | 28.81* (12.64)* | 28.32* (13.45)* | 26.67 (10.64) | 28.12* (10.12)* | 27.37* (10.35)* | 27.43 (10.66) | 26.16 (8.21) | 26.82 (9.48) |
| | Mean CATFLU | 15.67 (5.01) | 14.60 (4.80) | 15.18 (4.91) | NA NA | NA NA | NA NA | 17.90* (4.06)* | 17.46 (5.22) | 17.66* (4.70)* | NA NA | NA NA | NA NA | NA NA | NA NA | NA NA | NA NA | NA NA | NA NA | 14.52 (4.95) | 14.47 (5.24) | 14.50 (5.02) |

## 3.2 Methods

### 3.2.1 Development of composite score from cognitive outcomes using factor analyses

The multiple cognitive outcomes of our project, contained in the pooled database of Cognitive Training and Remediation Studies (see Table 3.2), were collected in order to assess Cognitive Remediation Therapy (CRT) effectiveness. Groups of outcomes were assumed to measure different aspects (memory, processing speed and executive function) of the same underlying construct: cognitive abilities. However, there was no reliable and validated scale summarising these outcomes. Because all outcomes were **commensurate outcomes**, i.e. measuring different aspects of the same underlying construct using the same scale (Teixeira-Pinto et al. 2009), a factor analysis (FA) will be used to find preferably one or few latent constructs able to explain the common variance of the outcomes and thus to summarize the outcomes as efficiently as possible. The estimated latent factor will then be used as dependent variable in a univariate MissForest-Lasso model. FA will be conducted in the following steps: exploratory factor analysis (EFA) followed by confirmatory factor analysis (CFA, used in the model development stage) will be run on the observed outcomes (called 'items' in FA), then I will confirm the measurement structure invariance across time using a longitudinal factor analysis (LFA, Meredith and Teresi 2006, see Paragraph below). Because most observed outcome variables were measured at three time points (baseline, end-of-treatment and follow-up) at least, the FA will use the three time points data. Even though I want to apply this FA results to a univariate regularised regression that only requires the end-of-treatment outcome (latent variable at end-of-treatment) as dependent variable and the baseline measure of the latent outcome as one of the covariates, I decided to also include the third time point to give more validity to the latent factor structure.

**Factor analysis of outcomes** The software used for the cross-sectional EFAs and CFAs and the LFA was Mplus7 (User's Guide: L. Muthén and B. Muthén 2012). The R package '`lavaan`' (Rosseel 2012) was used to produce unbiased estimates for the latent factor (see below the Paragraph 'Factor scores'). The analyses were restricted to the continuous items only, because methods computing unbiased estimates for the latent factor from mixed continuous and categorical indicators were not implemented in the softwares available for the project (B. Muthén and Yang Hsu 1993). Therefore, the six categorical items were excluded.

The default estimator for factor analysis of continuous data in Mplus7 was used: the Maximum Likelihood (ML) estimator, which assumes normality of the data. If there is excess kurtosis

in the continuous variables, the ML estimates will still be consistent when the model is correct. However, in this case standard errors and fit statistics will need correction (Lattin, Carroll, and P. Green 2003).

Firstly, the data were described and prepared for the FA: total scores variables (i.e. combinations of variable subgroups) among the outcomes were eliminated as well as one variable in pairs of highly correlated variables (correlation coefficient $\rho >$0.9). The *covariance coverage* between two variables is the proportion of observations that have values for both variables. When some items were measured for some studies and the other items were only available for the remaining studies, the covariance coverage between them was 0: i.e. no overlapping observations. One outcome variable in pairs of variables with 0 covariance coverage was excluded to initialise the Expectation Maximisation (EM) algorithm needed to run the analysis. The choice of these variables to be excluded was dictated by their relative frequency: the variable in the pair with 0 covariance coverage with more missing data was excluded. To meet the assumption of normality, very skewed variables (i.e absolute value of skewness greater than 1.5) were log-transformed if they had strictly positive values, or transformed through the Yeo-Johnson transformation (Yeo and R. Johnson 2000) if they also had non-positive values. The Expectation Maximisation algorithm (Dempster, Laird, and Rubin 1977, see Subsection 1.2.2) implemented in FA dealt with missing data in the factor model development process.

Secondly, **EFA** was used to explore the number of factors and assess the suitable items loading on the factors at the different time points (baseline, end-of-treatment and follow-up) separately, in order to identify a common factor structure. The EFA solutions corresponding to the Geomin oblique rotation (Lattin, Carroll, and P. Green 2003), which allows the factor to be correlated, were used. EFA models were done with different numbers of factors up to a maximum which allowed identifiability of the EFA model (e.g. two items cannot identify one factor, Lattin, Carroll, and P. Green 2003). EFA models were assessed according to:

1. the presence of Heywood cases, i.e.**negative residual variances** (primary indication of bad fit for the EFA). If an item shows negative residual variance, it means that it accounts for all or most of the variance of the latent factor and is therefore redundant. If there was a relative small number of items with negative residual variances, the corresponding items were excluded and the EFA was run again;

2. *Kaiser's eigenvalue-greater-than-one rule*: the number of eigenvalues (extracted variance in a factor)greater than 1 is considered as an estimate of the optimal number of common factors. However, this method of determining the number of factors sometimes yields an

unreasonably high number of factors (Costello and Osborne 2005), which would cause overfactoring (i.e. a large number of factors relative to the number of items);

3. the *scree plots*, which display the eigenvalues associated with a factor in descending order versus the number of the factor. The scree plot always shows a downward curve and the point prior to the 'elbow' (i.e. where the slope of the curve is clearly levelling off) indicates the number of factors that should be generated by the analysis. Scree plots are more efficient than Kaiser's rule, but they also have a tendency to overestimate the number of factors (Fabrigar et al. 1999).

4. the **variance explained** by the factors;

5. the absolute size of the **loadings** (if larger than or equal to 0.2 and significantly different from 0). A standardised factor loading is the correlation between observed item and latent variable. Items having weak loadings (i.e. smaller than 0.2 and non-significantly different from 0) on a factor were removed from the analysis and the EFA was rerun. The rotated cross-loadings determine the quality of the variables measuring the factors. Too many item cross-loadings are indication of a model without *simple structure* (i.e. the model structure given by the items loading significantly on only one factor), leading to poor interpretation. The rotation used for the factors to reach a simple structure was the oblique rotation Geomin;

6. the **interpretation** of the factors and their quality (number of variables loading on a factor, overfactoring (too many factors), factor determinacy if the items are all continuous and usefulness of factors);

7. the following **model fit assessments** (Lattin, Carroll, and P. Green 2003):

   - the *Chi-square test* statistic which tests that the implied model does not fit significantly worse than a model where the variables correlate freely (saturated model). P-values greater than 0.05 indicate good fit. However, this test statistic is sensitive to the sample size and thus less reliable than the other fit measures. Here is the formula of the Chi-square statistic:

     $$T = 2nF_{ML}(\hat{\pi}), \tag{3.1}$$

     where $n$ is the number of observations, $F_{ML}$ the fit function given by the difference between the log-likelihood for the saturated model and the implied model log-

likelihood and $\hat{\pi}$ is the maximum likelihood estimate under the null hypothesis, i.e. the implied model;

- the *Root Mean Square Error of Approximation* (RMSEA) which assesses the null hypothesis of approximate fit (rather than perfect fit). It is a function of the $\chi^2$ test statistic, the degrees of freedom of the implied model $d$ and the number of observations in the sample $n$:

$$RMSEA = \sqrt{\max\left(\frac{T^2 - d}{nd}, 0\right)}\sqrt{G}, \tag{3.2}$$

where $T$ is the $\chi^2$ test statistic as defined in (3.1) and $G$ the number of groups (in our case $G = 1$). Values of the statistic smaller than 0.05 indicate close fit. The RMSEA adjusts for sample size contrarily to the Chi-square test statistic.

EFA is an iterative process: when items poorly measured the factors or factors were poorly measured in the EFA, the EFA analysis was repeated by excluding one item (or factor) at a time. The items with negative residual variance had the priority to be excluded one at a time; then the items not loading significantly on a factor were excluded, by considering all the other assessments in combination, and the EFA was rerun after each item exclusion.

Thirdly, I tested the chosen number of factors and the factor structures through **Confirmatory Factor Analysis (CFA)** as suggested by the EFAs cross-sectionally. It is important to specify that here CFA is used in the model development stage and not to confirm a given factor structure supported by the literature. CFA can be used to develop a scale without necessarily confirming it on a sample different from the EFA to simply evaluate a scale's internal structure (Brown 2006). CFA models were assessed through the criteria below:

- Measures of goodness of fit (some of them were already used in the EFA):
  - the *Chi-square test* statistic (3.1);
  - the Root Mean Square Error of Approximation (*RMSEA*, (3.2));
  - the baseline comparison indexes: *Tucker and Lewis Index* (TLI) and *Comparative Fit Index* (CFI), that compare the fit of the baseline model with the implied model. The baseline model is the independence model, i.e the model with uncorrelated outcomes with unrestricted variances and unrestricted means and/or thresholds. Values of TLI and CFI less than 0.9 indicate poor fit; 0.9-0.95 indicate good fit, values close to 1 indicate very good fit and values grater than 1 might indicate overfitting. They

are on average the same across sample sizes and CFI has less variance than TLI.

$$TLI = \frac{\frac{T_b}{d_b} - \frac{T_m}{d_m}}{\frac{T_b}{d_b} - 1} \qquad CFI = 1 - \frac{\max(T_m - d_m, 0)}{\max(T_m - d_m, T_b - d_b, 0)}, \tag{3.3}$$

where $T_b$ and $T_m$ are the $\chi^2$ statistics for the baseline model (independence model) and the implied model respectively and $d_b$ and $d_m$ are the respective model degrees of freedom;

- The reproduced correlation matrix and the standardized residuals (discrepancies between observed and reproduced correlations, which measures is the *Standard Root Mean square Residual* fit index - SRMR, less than 0.08 for good fit - for normally distributed data) were considered;
- *Model modification indices* (estimated for all parameters that are fixed or constrained to be equal), i.e. the expected drop in chi-square if the parameter is estimated. Modification indices are considered to improve the fit of the model: for example it is good practice to free the constraint in the model that shows the highest modification index to ameliorate the model fit;
- *Factor loadings* for a simple structure;
- The percentage of *variance explained* by the factors in the items (the $R^2$);
- *Factors determinacy* which measures how well the scores for the latent factors are estimated. It is the correlation between the estimated score and the true score and ranges from 0 to 1 with 1 being best;
- *Factors discriminant validity* which is measured by the correlations between the factors and indicates factor uniqueness;

When competing CFA models were present, model selection was done according to the following criteria and tests:

- *non-nested models* were compared by looking at the following indices (if items were all continuous):
  - *Akaike's information criterion*, $AIC = T_m - 2d_m$; the model with the lower AIC is deemed the better fit;

  - *Bayesian information criterion*, $BIC = T_m - d_m \log n$; again choose the model with the smaller value. BIC rewards parsimony (fewer parameters), compared to AIC;

- *nested models* were compared through *Likelihood Ratio Test (LRT)* and AIC/BIC.

The three final cross-sectional CFA models (baseline, end-of-treatment and follow-up) were chosen by combining the assessment of the criteria above.

Finally, a **Longitudinal confirmatory Factor Analysis (LFA)** model was fit to the data to formally test invariant factorial structure across time. Factorial invariance is also known as measurement equivalence and metric invariance. The following longitudinal models were analysed in steps so that at each step constraints on parameters were added:

1. *Configural invariance (baseline model)*: the measurement model was assumed the same across time points without any constraint on loadings, intercepts/thresholds or variances;
2. *Metric (weak) invariance*: constraints on loadings were added so that they were equal across measurement occasions;
3. *Scalar (strong) invariance*: intercepts/thresholds of the items with same loading were constrained to be equal at each time point;
4. *Structural (full) invariance*:
   a) residual variances were constrained to be equal for non-invariant items,
   b) factor variances were constrained to be invariant across time,
   c) factor means were assumed the same across time.

The increasingly constrained nested models were compared with the LRT and when they had equivalent fit, the strictest (more parsimonious) model was chosen. When an invariance model did not test equal all the parameters supposed to be the same, then the parameters which could not be fixed across time were freed to be estimated and the model was called with the term 'partial invariance' instead of 'invariance' model. For example, if in a scalar invariance model one intercept could not be fixed to be equal to the others for model goodness of fit reasons, then that intercept was let free to be estimated and the returned model became a partial scalar invariance model.

**Factor scores** In order to obtain values for the single summary outcome latent measure to use as a dependent variable in the model, *factor scores* for the latent construct were estimated.

Factor scores are the estimated predictions for the unobserved values of the factors for individuals in the sample. To estimate the factor scores, the direction of the regression model is reversed: best prediction of the factors given the observed variables. Factor scores are weighted sums of the observed variables. Geometrically speaking, they are the locations of each of the individual observations in the reduced factor space (Lattin, Carroll, and P. Green 2003).

Factor scores for baseline and end-of-treatment were computed from cross-sectional 1-factor CFAs having the same factorial structure tested in LFA, which was invariant across time.

The factor scores were computed using the Bartlett method (Bartlett 1937 and Bartlett 1938), which is not implemented in Mplus7, but in R (Team 2016) with the package `lavaan` (functions: `cfa()`, with option `mimic="Mplus"` to have the same output as Mplus7 in the CFA; and `lavPredict()` to compute the factor scores, with option `method = "Bartlett"`). I did not use the factor scores directly from the LFA in order to avoid induced correlation between baseline and follow-up scores. The Bartlett method produces unbiased estimates of factor scores when factor indicators are continuous (SL 2005). Mplus7 only uses the regression method (Thomson 1934 and Thurston 1935), which computes biased estimates of factor scores that cannot be used as dependent variables (B. Muthén and Yang Hsu 1993 and Skrondal 2001). The Bartlett method estimates factor scores with mean 0 and variance being the squared multiple correlation between items and factor. Therefore, factor scores were standardised to variance 1.

Factor scores estimation allows to analyse the units with at least one observed item assuming missingness is MAR. This allows to estimate factor scores for patients with one or more unmeasured items.

**Meta-analysis of factor scores** A meta-analysis of factor scores was run in order to see if the factor scores showed a CRT positive effect in the seven DoCTRS RCTs (see Subsection 3.1.1) as well as to assess between study heterogeneity. First, I estimated factor scores treatment effect sizes for each study. Then, I collapsed the results by performing a meta-analysis. Because I assumed between study heterogeneity, a random effects meta-analysis model was used (Armitage, Berry, and Matthews 2002). Effect sizes (Cohen's $d_i, i = 1, \ldots, 7$) for the seven studies were calculated as follows:

$$d_i = (m_i^t - m_i^c)/SD_i^{\text{pooled}}, \tag{3.4}$$

where $m_i^t$ and $m_i^c$ are the means for the cases and the control respectively for the study $i, i = 1, \ldots, 7$, and $SD_i^{\text{pooled}}$ indicates the pooled standard deviation for the two groups for study $i$.

The standard error ($SE_i$) of the effect size was estimated using the following formula:

$$SE = \sqrt{N_i/(n_i^t n_i^c) + d_i^2/2(N_i - 2)}, \tag{3.5}$$

where $n_i^t$ and $n_i^c$ are the numbers of patients in the treatment and the control group respectively for the study $i$, and $N_i$ $(= n_i^t + n_i^c)$ was the sample size of study $i$.

I also explored the degree of heterogeneity across the studies using the $Q$ and the $I^2$ statistic. The $Q$ statistic is based on the squared differences between each study effect and their fixed effect average, and $I^2$ is derived from $Q$ and is interpreted as the proportion of total variability explained by the heterogeneity between studies. If the studies are homogeneous both should be small.

The random effects meta-analysis with the above effect size formulas and heterogeneity tests is implemented in the command `metan` (Bradburn, Deeks, and D. Altman 1998) of the statistical software Stata 14 (Stata 2015).

**Factor scores for 'Fiszdon'** The data from the study 'Fiszdon 1' (see Subsection 3.1.1) were used as validation set for the prediction models to develop, thus factor scores for this study needed to be computed. The 23 subjects in common between the studies 'Fiszdon 1' and 'Fiszdon 2' also had values for other outcomes measured in 'Fiszdon 2' and not in 'Fiszdon 1'. Therefore, I also considered the 23 subjects' data from 'Fiszdon 2' in order to compute factor scores for the external validation of the prediction models. It is valid to use variables from another RCT measured on common participants to compute factor scores when the underlying FA model is longitudinal and tested for measurement invariance across time. This property is

called *group invariance of item parameters* (Baker 2001): if the estimated LFA has a good fit, then the parameters do not change across time, independently of when the score was taken from the same people, provided that it is a pre and post-treatment score and with the same treatment.

### 3.2.2 MissForest-Lasso precision medicine models

Three MissForest-Lasso precision medicine models (see Subsection 2.2) were developed:

1. **Model 1** having the factor scores as outcome (see Subsection 3.2.1;
2. **Model 2a** having the clinically important cognitive measure of executive function Wisconsin Card Sorting Test Perseverative Errors (WCST PE); only the observations with the outcome were included;
3. **Model 2b** having one of the most important cognitive measure with least missing data as outcome; all the observations belonging to the studies that measured the outcome (also observations with missing outcome) were included.

The first model was built because there is an interest in analysing multiple cognitive outcomes simultaneously with a summary measure of memory, processing speed and executive function as a dependent variable. Only the observations with estimated factor scores were included in the analysis as the FA already included the uncertainty due to a missing data imputation process: the EM algorithm.

In contrast, Models 2a/b were developed in order to apply MissForest-Lasso to an observed outcome which has already been validated and is known to be reliable unlike the factor scores, and to compare performances with complete and missing outcome. In Model 2b, missing data for WCST PE will be imputed by MissForest and not in a factor analysis procedure.

To develop the models, I followed the general guidelines by Steyerberg and Vergouwe (2014), which consists of seven phases described in Figure 1.1 (see Subsection 1.2.1. In the different steps, I did the following:

- **Step 1 (problem definition)**: aims of the prediction problem were defined as well as the outcome of interest. Summary statistics for the outcomes were presented per treatment arm (CRT vs TAU).
- **Step 2 (coding of predictors)**: categorical predictors were transformed into dummies (binary factors contrasting the reference level and another level of the categorical variable,

for example if a categorical variable has 5 categories, then it will generate 4 dummies as covariates in the model). Categorical variables with small number of observations within a level ($\leq$ 10) were recoded in order to avoid near-zero-variance dummy predictors (Kuhn and K. Johnson 2013). Furthermore, the levels of categorical variables were coded so that the reference level was the one with most observations. The ordinal covariates were treated as continuous.

- **Step 3 (model specification)**: although fully pre-specified prediction models (i.e. models including all available variables) minimise selection bias (Steyerberg 2009), our model was specified by dropping the variables with more than 70% missing data to reduce the bias due to imputation. If there was perfect multicollinearity between variables (correlation coefficient 1 or -1), the most useful variable in each group of perfectly correlated variables was chosen based on clinical expertise. Categorical variables with more than seven levels, which could not be collapsed in a smaller number of levels, were dropped for model interpretability due to small sample size. In Model 1, both factor scores at baseline and single observed baseline items, from which factor scores were computed, were included in the model linear predictor. This allows identifying the key cognitive variables as predictors and may increase prediction accuracy. This is feasible as the Lasso allows including correlated variables without resulting in multicollinearity, by selecting only one variable among strongly correlated variables (Zou and Hastie 2005). Also different scales of PANSS variables were included to assess which scale had more predictive power. The study ID variable was not included in the model. Instead the study information variables were used in order to make the model generalisable to new study data.

- **Step 4 (model estimation)**: MissForest missing data imputation computed random forest models with the default settings: 100 trees for forest, the maximum number of iterations to be performed (given that the stopping criterion is not met beforehand) being 10, and the default number of predictors sampled for splitting at each node of a tree being $\max\left(\lfloor p/3 \rfloor, 1\right)$ for continuous dependent variables and $\left\lfloor \sqrt{p} \right\rfloor$ for categorical ones, where $p$ is the number of variables in the imputation model (see section 2.1.2 for details about MissForest). After imputing the data through MissForest, the Lasso models on the completed data were estimated with bootstrap tuning (100 bootstrap samples and 40 values for the tuning parameter $\lambda$, see Subsection 2.2). The grid for $\lambda$ was decided in order to avoid the Lasso choice of the null model (intercept only), which is of no interest to us. This would result in constant-valued predictions, which have zero variance and undefined

pseudo-$R^2$. To define such a range for $\lambda$, it is sufficient to run the model once without specified grid and look at the resampling performances in the output: the tuning parameters that correspond to missing data in the estimates of pseudo-$R^2$ are not useful (these are usually large). Simulations (see Section 2.3) suggested that the 3% tolerance $\lambda$ model (delivering an MSE within 3% of the minimum) provides the best compromise between variable selection and prediction accuracy. Also the models corresponding to the following $\lambda$s were estimated: the best $\lambda$, the one-SE tolerance $\lambda$ and the 15% tolerance $\lambda$ as in the simulation study (see Subsection 2.2), in order to again confirm the 3% tolerance model as the model at the same time interpretable and good in prediction accuracy.

- **Step 5 (model performance)**: estimates of apparent discrimination and calibration performance for MissForest-Lasso were presented (for definitions, see Subsection 1.2.1). First, MissForest imputation accuracy was assessed with the out-of-bag (OOB) imputation normalized root mean squared error (NRMSE, see Oba et al. 2003) for the continuous part of the imputed data, and with the proportion of falsely classified entries (PFC) for the categorical part of the imputed data set (in both cases good performance of MissForest leads to a value close to 0 and bad performance to a value around 1). Second, Lasso apparent discriminative performance was evaluated with the MSE and the pseudo-$R^2$ by applying the model on the development data; apparent calibration was measured through the calibration slope $\beta$ and the calibration-in-the-large $\alpha$ (see Subsection 1.2.1). Predictions-versus-observed values plots, corresponding to the different $\lambda$s models, were compared.

- **Step 6 (model validity)**: The MissForest-Lasso models were internally validated using the bootstrap optimism-correction as for Harrell, Lee, and Mark 1996 (see Subsection 2.1.3) as describe in the simulations methods (see Section 2.2). Apparent performance estimates were corrected with the bootstrap-estimated optimism to obtain internally validated performance estimates. Internal and external optimisms were compared (see Table 2.1 for definitions).

- **Step 7 (model presentation)**: the three chosen models were the models corresponding to the 3% tolerance penalty $\lambda$ according to the simulation results with 100 covariates (see Subsection 2.3.2). The chosen models' linear predictors were recalibrated with the optimism-corrected calibration parameters $\beta_{corr}$ and $\alpha_{corr}$ (see Subsection 1.2.1). Let us write the model as $M(\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{b}$, where $\boldsymbol{X}$ is the matrix of predictors plus a vector identically equal to 1 for the intercept and $\boldsymbol{b}$ is the vector of the estimated co-

efficients. Thus, the recalibrated model $M_{recalibrated}$ was obtained through the formula: $M_{recalibrated}(\boldsymbol{X}) = \alpha_{corr} + \beta_{corr}\boldsymbol{Xb}$, and recalibrated coefficients were presented in a table.

Finally, I compared the three models in terms of prediction accuracy and variable selection.

### 3.2.3 Secondary analysis: MissForest-Lasso prognostic models

Models 1, 2a and 2b described in Subsection 3.2.2 were rerun with the same method without assuming moderation of treatment, i.e. without including interaction terms in the linear predictor, to obtain Models 3, 4a and 4b respectively. This secondary analysis was done to assess the importance of eventually selected moderators, by looking at the models' changes in prediction accuracy and variable selection.

## 3.3 Results

### 3.3.1 Development of composite score from cognitive outcomes using factor analyses: results

**Data preparation** The 32 continuous outcomes measuring cognition (12 memory, one processing speed and 19 executive function variables) were available for 467 out of 468 patients (see Table 3.2).

At least 19 outcome variables were measured at three time points (baseline, end-of-treatment and follow-up), the others were only present at baseline and end-of-treatment (included some mid-point measurement which was not used in the analysis).

The outcomes were approximately normally distributed based on histogram, skewness and kurtosis assessments apart from two outcomes. These two outcomes were exaggeratedly positively skewed and leptokurtic at baseline and end-of-treatment: TMTA and TMTB (see Table 3.2 for definitions; kurtosis was 75.0 and 10.0 for TMTA at baseline and end-of-treatment respectively, 19.7 and 65.3 for TMTB; skewness was 2.6 and 7.0 for TMTA, 3.8 and 6.6 for TMTB). Both variables were log-transformed and skewness and kurtosis were reduced to acceptable levels: skewness now was less than 1 and kurtosis less than 2.5. Furthermore, to improve normality in the variable CATFLU, one subject was removed from the data (ID 145) because he/she had CATFLU=58 at follow-up, meaning that a patient could name 58 animals in 60 seconds, which is unlikely. In agreement with the PhD project clinician, I considered it to be a typographical error.

Table 3.4: Baseline, end-of-treatment and follow-up **correlation matrices for the outcomes** that are not total scores and have pairs positive covariance coverage: the correlation between each pair of variables is computed using all complete pairs of observations on those variables.

| | LNS | WAIS D | WAIS PC | CATFLU | FAS A | FAS NR | TMTA | TMTB | WCST NE | WCST PC | WCST PE |
|---|---|---|---|---|---|---|---|---|---|---|---|
| LNS | | | | | | | | | | | |
| WAIS D | 0.406 | | | | | | | | | | |
| | 0.508 | | | | | | | | | | |
| | 0.627 | | | | | | | | | | |
| WAIS PC | 0.552 | 0.325 | | | | | | | | | |
| | 0.407 | 0.458 | | | | | | | | | |
| | 0.465 | 0.351 | | | | | | | | | |
| CATFLU | 0.492 | 0.280 | 0.308 | | | | | | | | |
| | 0.474 | 0.284 | 0.271 | | | | | | | | |
| | 0.034 | 0.109 | 0.176 | | | | | | | | |
| FAS A | 0.527 | 0.373 | 0.292 | 0.424 | | | | | | | |
| | 0.384 | 0.381 | 0.176 | 0.601 | | | | | | | |
| | 0.485 | 0.449 | 0.280 | 0.044 | | | | | | | |
| FAS NR | 0.584 | 0.379 | 0.272 | 0.462 | 0.981 | | | | | | |
| | 0.453 | 0.423 | 0.161 | 0.449 | 0.982 | | | | | | |
| | 0.467 | 0.483 | 0.284 | 0.332 | 0.979 | | | | | | |
| TMTA | −0.382 | −0.146 | −0.473 | −0.451 | −0.524 | −0.527 | | | | | |
| | −0.445 | −0.133 | −0.334 | −0.439 | −0.326 | −0.368 | | | | | |
| | −0.363 | −0.268 | −0.423 | −0.255 | −0.378 | −0.367 | | | | | |
| TMTB | −0.476 | −0.055 | − 0.434 | −0.303 | −0.282 | −0.365 | 0.620 | | | | |
| | −0.364 | −0.061 | −0.404 | −0.531 | −0.160 | −0.262 | 0.598 | | | | |
| | −0.473 | −0.355 | −0.615 | −0.282 | −0.269 | −0.343 | 0.575 | | | | |
| WCST NE | −0.277 | −0.040 | −0.177 | −0.010 | −0.073 | −0.151 | 0.188 | 0.106 | | | |
| | −0.248 | 0.018 | 0.006 | −0.192 | 0.025 | −0.069 | 0.052 | 0.040 | | | |
| | −0.063 | −0.087 | −0.104 | −0.131 | 0.060 | −0.050 | 0.046 | 0.060 | | | |
| WCST PC | 0.411 | 0.020 | 0.478 | 0.240 | 0.113 | 0.284 | −0.241 | −0.346 | −0.476 | | |
| | 0.471 | 0.032 | 0.243 | 0.319 | 0.089 | 0.161 | −0.304 | −0.348 | −0.549 | | |
| | 0.293 | 0.208 | 0.362 | 0.220 | 0.103 | 0.208 | −0.247 | −0.313 | −0.104 | | |
| WCST PE | −0.245 | −0.084 | −0.429 | −0.347 | −0.073 | −0.210 | 0.165 | 0.372 | −0.040 | −0.726 | |
| | −0.344 | −0.059 | −0.180 | −0.301 | −0.177 | −0.282 | 0.380 | 0.488 | 0.070 | −0.734 | |
| | −0.309 | −0.207 | −0.335 | −0.214 | −0.076 | −0.127 | 0.283 | 0.337 | 0.051 | −0.677 | |

Four continuous outcome variables were total scores of subgroups of outcomes: CVLT TR (summarising the other two CVLT outcomes, see Table 3.2), HAY T and HAY TE (combining the two HAY tests sections A and B), WCST TE (combining the two WCST error measures, see Table 3.2). These total scores were excluded from the FA.

After excluding the total scores variables, 21 outcome variables had 0 covariance coverage with the other variables (no observations in common). As there was no preference of retaining any variables from a clinical perspective, the outcomes with higher percentage of missing data that had 0 covariance coverage with most other variables were eliminated, so that only 11 outcomes were left. The latter had covariance coverage in the range 0.087-0.892 at baseline, 0.075-0.872 at the end-of-treatment and 0.107-0.966 at follow-up.

Available data per time point for the 11 items were: 461 patients at baseline, 413 at the end-of-treatment and 291 at follow-up.

More than 2/3 of the correlations between the 11 items (see Table 3.4) looked stable or got weaker across the 3 time points, as expected. Among the variables, two were highly correlated: the correlation coefficients between FAS A and FAS N were 0.981 at baseline, 0.982 at end-of-treatment and 0.979 at follow-up (see Table 3.4). Therefore, the item FAS A with higher percentage of missing data was excluded.

Therefore, only ten items were used to conduct the FA: two items were memory outcomes, one was a processing speed outcome and the other eight were executive function outcomes:

- Memory outcomes:
  - LNS: Letter-Number Span (n of correct trials), measured in ‘Keefe’, ‘Keshavan’ and ‘Wykes 3’;
  - WAIS D: Wechsler Adult Intelligence Scale Digit Span, measured in all studies apart from ‘Keefe’

- Processing speed:
  - TMTA: Trailmaking test part A (Paper & pencil), time to completion (seconds), measured in ‘Bell’, ‘Keefe’, ‘Wykes 2’ and ‘Wykes 3’;

- Executive function outcomes:
  - CATFLU: Category fluency, Animal naming (n animals named in 60 s); measured in ‘Bell’, ‘Keefe’ and ‘Wykes 3’;

- FAS NR: Verbal fluency (FAS), Total number of correct responses, measured in 'Wykes 1', 'Wykes 2', 'Wykes 3' and 'Bell';
- TMTB: Trailmaking test Part B (Paper & pencil), time to completion (seconds), measured in 'Bell', 'Wykes 2', 'Wykes 3' and 'Keshavan';
- WAIS PC: Wechsler Adult Intelligence Scale Picture Completion, measured in 'Wykes 1', 'Bell' and 'Wykes 3';
- WCST NE: Wisconsin Card Sorting Test, Non-Perseverative Errors (0 to 128), measured in 'Wykes 1', 'Bell', 'Wykes 2', 'Wykes 3' and 'Keshavan';
- WCST PC: Wisconsin Card Sorting Test, Percent Conceptual Responses (0 to 100), measured in 'Wykes 1', 'Bell', 'Wykes 2', 'Wykes 3' and 'Keshavan';
- WCST PE: Wisconsin Card Sorting Test, Perseverative Errors (0 to 128), measured in 'Wykes 1', 'Bell', 'Wykes 2', 'Wykes 3', 'Keshavan' and 'Circuits'.

Their mean percentage of missing data (including missingness by design) was 55.4% (46.9% at baseline, 54.5% at the end-of-treatment and 64.9% at follow-up).

**Outcomes Cross-Sectional Exploratory Factor Analyses (EFA) at baseline, end-of-treatment and follow-up**

**Baseline EFA** The EFA at baseline used 460 out of 467 observations as at baseline the 10 outcomes had values for only 460 patients. The item WCST PC was removed because its negative residual variance prevented the 2 and 3-factor EFAs from achieving convergence.

| Factors | Parameters | Chi-square | | | RMSEA | Negative res. variance | Weak loadings |
|---|---|---|---|---|---|---|---|
| | | $\chi^2$ | df | p-value | | | |
| 1 | 27 | 90.298 | 27 | $<$0.0001 | 0.071 | - | - |
| 2 | 35 | 46.474 | 19 | 0.0004 | 0.056 | - | WCST NE |
| 3 | 42 | 20.093 | 12 | 0.0653 | 0.038 | WAIS PC, WCST NE | WCST PE |

Table 3.5: EFA of 9 items, fit measures at baseline (n=460). Abbreviations: res=residual

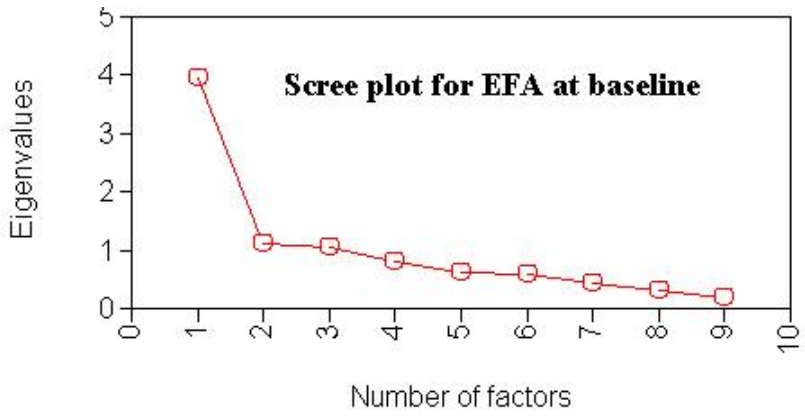
The EFA RMSEA and $\chi^2$-test in Table 3.5 suggest that the factor model with three factors presents the best fit. Kaiser's rule of eigenvalues greater than one suggested the existence of three factors (the first eigenvalue was 4.0 while the rest were less than 1.2) and this somewhat confirmed the EFA RMSEA and $\chi^2$ test results. Instead, the scree plot suggested 2 factors for these data (see Figure 3.1). Moreover, the 3-factor structure estimated a negative residual

variance for the item WCST PE and removing it leads to overfactoring. Therefore, due to the criteria described in the methods, I considered only the 1 and 2-factor structures in Table 3.5, which would fit the data better:

- 1 FACTOR: the 1-factor model (9 items) did not yield any negative residual variances or weak loadings and showed an almost good RMSEA (0.071) and 38% of variance explained.
- 2 FACTORS: WCST NE was dropped from the analysis because of its weak loading and the EFA was rerun. The analysis showed very good fit (RMSEA=0.032, variance explained=43%) and the factor structure was the following:
  - 1$^{st}$ factor: WAIS D, LNS, CATFLU, FAS N, WAIS PC
  - 2$^{nd}$ factor: WAIS D, TMTA, TMTB, WCST PE.

  The cross-loading WAIS D was much stronger on the first factor (0.826 on the first factor vs 0.526 on the second). However, the two factors were highly correlated ($\rho$ =-0.738).

Figure 3.1: Scree plot for EFA at baseline, nine continuous outcomes, 460 observations




**End-of-treatment EFA** This analysis used 412 out of 467 observations as the ten outcomes had values for only 412 patients at the end-of-treatment. In order to make the EFA converge for two factors, the item WCST PC was excluded (the 3-factor analysis did not converge even after eliminating this item).

The scree plot was similar to the one from the EFA at baseline (see Figure 3.1), suggesting 2 factors, and again three eigenvalues were larger than one, with the first being much larger than the others.

| Factors | Parameters | Chi-square | | | RMSEA | Negative res. variance | Weak loadings |
|---|---|---|---|---|---|---|---|
| | | $\chi^2$ | df | p-value | | | |
| 1 | 27 | 96.715 | 27 | <0.0001 | 0.079 | - | - |
| 2 | 35 | 34.025 | 19 | 0.0183 | 0.044 | WAIS D | - |
| 3 | N/A | N/A | N/A | N/A | N/A | N/A | N/A |

Table 3.6: EFA of nine items, fit measures at the end-of-treatment (n=412). N/A stands for 'not available' because of non-convergence. Abbreviations: res=residual

- 1 FACTOR: the 1-factor model (9 items) did not show any negative residual variances or weak loadings. The RMSEA was 0.079 and the variance explained was 37%.
- 2 FACTORS: By removing the item with negative residual variance WAIS D (Table 3.6), the model with two factors did not converge.

**Follow-up EFA** Because of loss to follow-up, only 290 out of 467 observations were available for the ten items at follow-up. Again the outcome WCST PC was dropped from the EFA analysis in order to achieve convergence for two factors (the 3-factor analysis did not converge even after eliminating this item). The scree plot continued to suggest 1 factor also at follow-up and the number of eigenvalues higher than one was two.

| Factors | Parameters | Chi-square | | | RMSEA | Negative res. variance | Weak loadings |
|---|---|---|---|---|---|---|---|
| | | $\chi^2$ | df | p-value | | | |
| 1 | 27 | 71.662 | 27 | <0.0001 | 0.076 | - | WCST NE |
| 2 | 35 | 30.469 | 19 | 0.0461 | 0.046 | - | WCST NE |
| 3 | N/A | N/A | N/A | N/A | N/A | N/A | N/A |

Table 3.7: EFA of nine items, fit measures at follow-up (n=290). N/A stands for 'not available' because of non-convergence. Abbreviations: res=residual

- 1 FACTOR: By removing the item WCST NE with non-significant loading (see Table 3.7), the EFA did not converge because of the negative residual variance of the item WAIS PC. After dropping this further outcome, the analysis converged without any problems with a poor RMSEA = 0.107 (7 items) but a better amount of variance explained: 42% compared to the other time points.
- 2 FACTORS: Similarly to the 1-factor case, WCST NE and WAIS PC were removed and the model with two factors showed two weak loadings. Excluding this two further items (one at a time) will lead to over-factoring: five items loading on only two factors.

**EFA conclusion** In conclusion, two or three factor solutions did not provide meaningful solutions at all three time points. The 1-factor model was possible at all time points. The 1-factor structure was similar across time: it was based on the same 9 items at baseline and end-of-treatment and 7 of the 9 items at follow-up. The 1-factor model was therefore selected. Also, the 1-factor structure makes sense from a clinical perspective as different cognitive outcomes can be summarised with one latent construct. The single factor model explained 38%, 37% and 42% of the variance at baseline, end-of-treatment and follow-up respectively.

**Cross-Sectional Confirmatory Factor Analyses at baseline, follow-up and end-of-treatment**

In the next step, I performed a CFA to further validate/confirm the 1-factor model at each time point. According to the EFA identified structures, in the CFAs at baseline and end-of treatment the 9-item structure was confirmed, while at follow-up the 7-item structure was analysed.

**Baseline CFA, N=460** By testing the 9-item structure of the EFA at baseline, the CFA model obtained showed almost good fit (RMSEA=0.071, 90% confidence interval (CI): 0.055–0.088; CFI=0.873; TLI=0.831; SRMR=0.071). However, there was no evidence that the factor explained some variance in the item WAIS D (variance explained ($R^2$)=0.032, $p$-value=0.139) and there was only some evidence that item WCST NE was explained by the factor ($R^2$=0.066, $p$-value=0.036).

**End-of-treatment CFA** The 9-item end-of-treatment configuration fit the data similarly to the baseline CFA (RMSEA=0.070, 90% CI: 0.054–0.087; CFI=0.867; TLI=0.822; SRMR= 0.093). Like in the baseline CFA, the $R^2$ for item WAIS D was not significantly different from 0 ($R^2$=0.024, $p$-value=0.228) as well as the $R^2$ for item WCST NE ($R^2$=0.061, $p$-value=0.054).

**Follow-up CFA** The CFA testing the 7-item structure at follow-up converged with inadequate fit: RMSEA=0.107, 90% CI: 0.081–0.136; CFI=0.839; TLI=0.758; SRMR=0.073.

**CFA conclusion** In summary, the CFAs showed only fair fit (at baseline and end-of-treatment) or inadequate fit (at follow-up) based on fit indexes. However, the estimated factor loadings were similar across time points apart from the loading for the item WAIS D, which was much stronger at follow-up compared to baseline and end-of-treatment (see Table 3.8). The confirmed item structures at baseline and end-of-treatment included the two memory outcomes

| Items | Baseline (N=460) | | End-of-treatment (N=412) | | Follow-up (N=290) | |
|---|---|---|---|---|---|---|
| | Estimate (SE) | P-value | Estimate (SE) | P-value | Estimate (SE) | P-value |
| LNS | 0.771 (0.041) | <0.001 | 0.698 (0.046) | <0.001 | 0.790 (0.059) | <0.001 |
| WAIS D | 0.178 (0.060) | 0.003 | 0.156 (0.065) | 0.016 | 0.622 (0.059) | <0.001 |
| CATFLU | 0.635 (0.058) | <0.001 | 0.699 (0.057) | <0.001 | 0.635 (0.086) | <0.001 |
| FAS N | 0.644 (0.051) | <0.001 | 0.546 (0.061) | <0.001 | 0.645 (0.058) | <0.001 |
| TMTA | -0.765 (0.035) | <0.001 | -0.776 (0.034) | <0.001 | -0.653 (0.064) | <0.001 |
| TMTB | -0.885 (0.027) | <0.001 | -0.886 (0.029) | <0.001 | -0.719 (0.059) | <0.001 |
| WCST PE | -0.477 (0.049) | <0.001 | -0.583 (0.048) | <0.001 | -0.373 (0.067) | <0.001 |
| WCST NE | -0.258 (0.062) | <0.001 | -0.247 (0.064) | <0.001 | - | - |
| WAIS PC | 0.593 (0.069) | <0.001 | 0.615 (0.077) | <0.001 | - | - |

Table 3.8: 1-factor cross-sectional CFA based on EFA (n=460 at baseline with nine items, n=412 at the end-of-treatment with nine items, and n=290 at follow-up with seven items): standardised parameter estimates (factor variance=1)

(LNS and WAIS D), one processing speed outcome (TMTA) and six executive function outcomes (CATFLU, FAS N, TMTA, TMTB, WAIS PC, WCST NE and WCST PE), while at follow-up it included the two memory items, the processing speed item and only four of the executive function outcomes (CATFLU, FAS N, TMTB and WCST PE). I am confident that the afore mentioned followed procedures and presented final models constitute reasonable measurements for memory, processing speed and executive function.

### Longitudinal Confirmatory Factor Analysis

The cross-sectional results suggested that different factor structures may be necessary at baseline/end-of treatment and follow-up. I explored this further by examining the impact of forcing a common pattern on the participants with data in at least one included item at all time points. This was done by applying the following configuration of six items to the single longitudinal factor:

- memory outcome: LNS
- processing speed: TMTA,
- executive function outcomes: CATFLU, FAS N, TMTB and WCST PE

The items WAIS D and WCST NE were excluded because their variance was not explained by the factor in the EFAs. The item WAIS PC was excluded because prevented the missing data algorithm in the LFA from converging.

The model with fixed structure across time, exhibiting configural invariance, provided a very good fit (RMSEA = 0.034, 90% confidence interval (CI): 0.024–0.044; CFI = 0.974 and SRMR = 0.070). The fit was equally good when factor loading invariance was imposed (p = 0.793 from

the $\chi^2$-test for difference testing between configural and metric invariance models - LRT) which suggests that a similar meaning to the latent constructs under study applies at baseline, end-of treatment and follow-up (RMSEA = 0.032, 90% CI: 0.021–0.041; CFI = 0.975 and SRMR = 0.079). The model that constrained also the means to be equal across time (scalar invariance) did not fit significantly differently from the metric invariance model and showed very good model fit (p = 0.333 for LRT, RMSEA = 0.031, 90% CI: 0.021-0.040; CFI = 0.975 and SRMR = 0.076). Structural invariance was explored by constraining residual variances to be the same across time as well as factor covariances and means. The residual variance invariance model was not significantly different from the scalar invariance model (p=0.502 for LRT), but modification indices suggested to free the residual variance for the item log(TMTB) at follow-up (modification index=9.714 and expected parameter change=0.044) to have a better fit. After doing so, the obtained partial residual variance invariance model showed a good fit, equivalent to the other more flexible invariance models, with RMSEA = 0.031, 90% CI: 0.021-0.040; CFI = 0.973 and SRMR = 0.083. Factorial covariance and factor mean invariance were next investigated and a final partial factor covariance and mean invariance model was obtained: RMSEA = 0.030, 90% CI: 0.020-0.039; CFI = 0.974 and SRMR = 0.083, equivalent in fit to the partial residual variance invariance model: LRT p=0.969 (see Table 3.9).

Because all the invariance models in Table 3.9 showed equivalent fit, the strictest (most parsimonious) model was chosen, i.e. the partial structural (factor covariance and mean) invariance model with the following goodness of fit indexes:

- RMSEA: 0.030 (90% confidence interval 0.020-0.039)
- CFI: 0.974
- TLI: 0.973
- SRMR: 0.083

Table 3.9: Longitudinal confirmatory factor (LFA) analsysis with six continuous outcomes (463 observations). Abbreviations: LL=loglikelihood, LRT=likelihood ratio test, P=partial, res. var.= residual variance, Cov=covariance. LRT tests for the invariance models are against the less constrained model above

| **Invariance** | **Constraints across time** | **LL (df)** | **LRT** | **RMSEA (90% CI)** | **CFI** | **SRMR** |
|---|---|---|---|---|---|---|
| Configural | - | -9109.729(75) | - | 0.034(0.024-0.044) | 0.974 | 0.070 |
| Metric | loadings | -9112.857(65) | p=0.793 | 0.032(0.021-0.041) | 0.916 | 0.079 |
| Scalar | intercepts | -9118.517(55) | p=0.333 | 0.031(0.021-0.040) | 0.975 | 0.076 |
| P. res. var. | all res. var. but TMTB | -9125.521(45) | p=0.173 | 0.031(0.021-0.040) | 0.973 | 0.083 |
| P. factor cov. and mean | factor cov. and mean but 1 cov. | -9125.552(43) | p=0.969 | 0.030(0.020-0.039) | 0.974 | 0.083 |

The estimated loadings are all above 0.4 shown in Table 3.10.

| **Items** | **Estimate (SE)** | | | | **p-value** |
|---|---|---|---|---|---|
| | Unstandardised | Standardised | | | |
| | (same across time) | baseline | end-of-treat | follow-up | |
| LNS | 2.887 (0.214) | 0.732 (0.034) | 0.739 (0.034) | 0.728 (0.037) | $<$0.001 |
| CATFLU | 3.374 (0.335) | 0.659 (0.044) | 0.667 (0.043) | 0.655 (0.046) | $<$0.001 |
| FAS N | 7.077 (0.693) | 0.601 (0.045) | 0.609 (0.045) | 0.597 (0.047) | $<$0.001 |
| log(TMTA) | -0.382 (0.024) | -0.773 (0.026) | -0.780 (0.025) | -0.770 (0.027) | $<$0.001 |
| log(TMTB) | -0.550 (0.030) | -0.892 (0.021) | -0.896 (0.021) | -0.823 (0.031) | $<$0.001 |
| WCST PE | -10.015 (0.986) | -0.480 (0.041) | -0.488 (0.041) | -0.476 (0.041) | $<$0.001 |
| | **Baseline** | **end-of-treatment** | **follow-up** | | |
| **Factor variance** | 1.000 ( - ) | 1.041* (0.039) | 0.977* (0.082) | | |
| **Factor means** | 0.000 ( - ) | 0.297* (0.032) | 0.297* (0.032) | | |

Table 3.10: Partial structural invariance model parameter estimates (n=463, six items, one factor at three time points). The unstandardised estimates were the same across time. The star sign * indicates $<$0.001

**LFA Conclusion** The 1-factor LFA analysis on six continuous items (measuring memory, processing speed and executive function: LNS, CATFLU, FAS, log(TMTA), log(TMTB) and WCST PE) and three time points (baseline, end-of-treatment and follow-up, see Figure 3.2 and Table 3.10) tested partial structural invariance showing a good fit (see Table 3.9). This suggested that the structural model was a reliable measure of the underlying cognition construct.

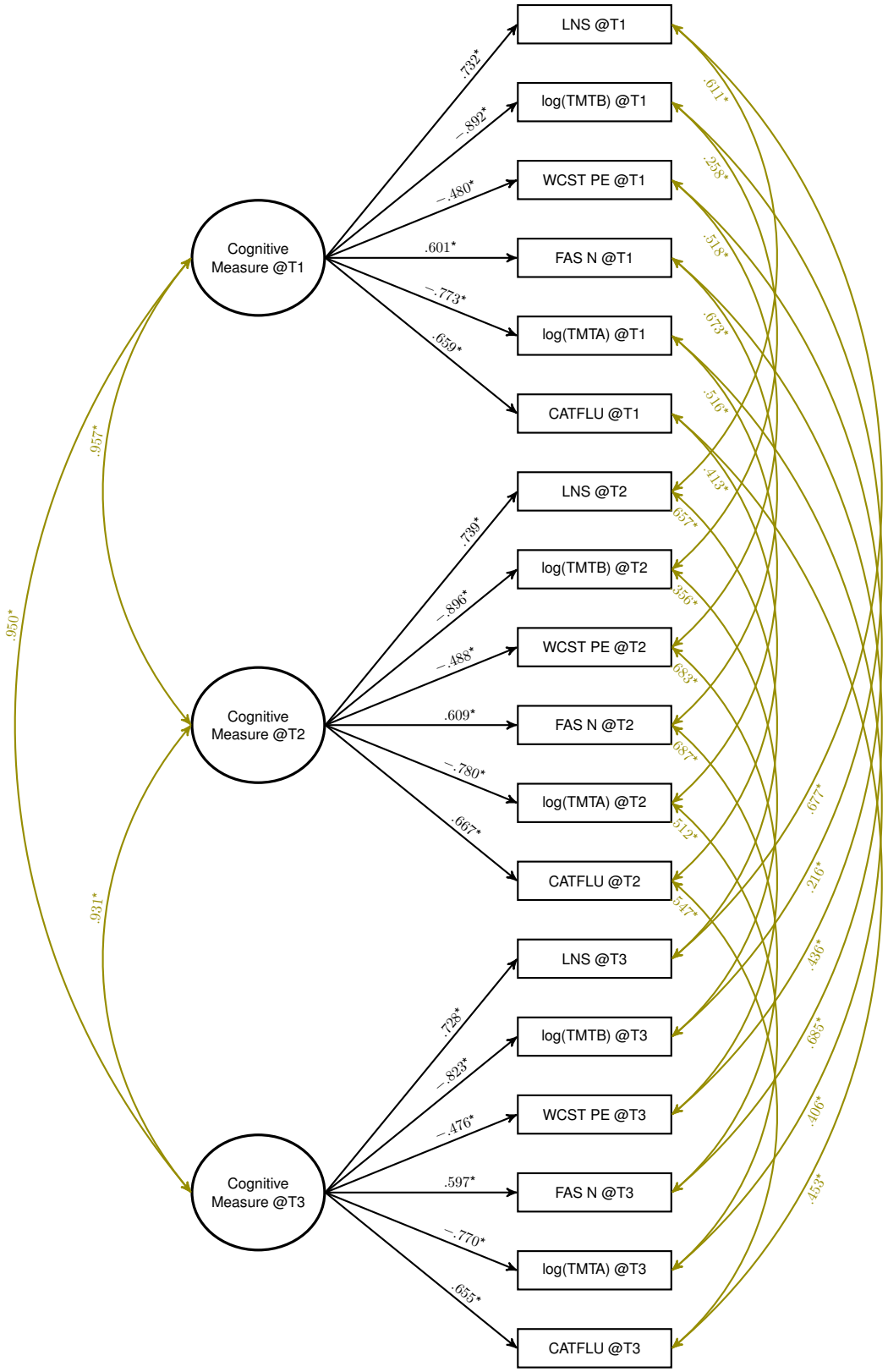
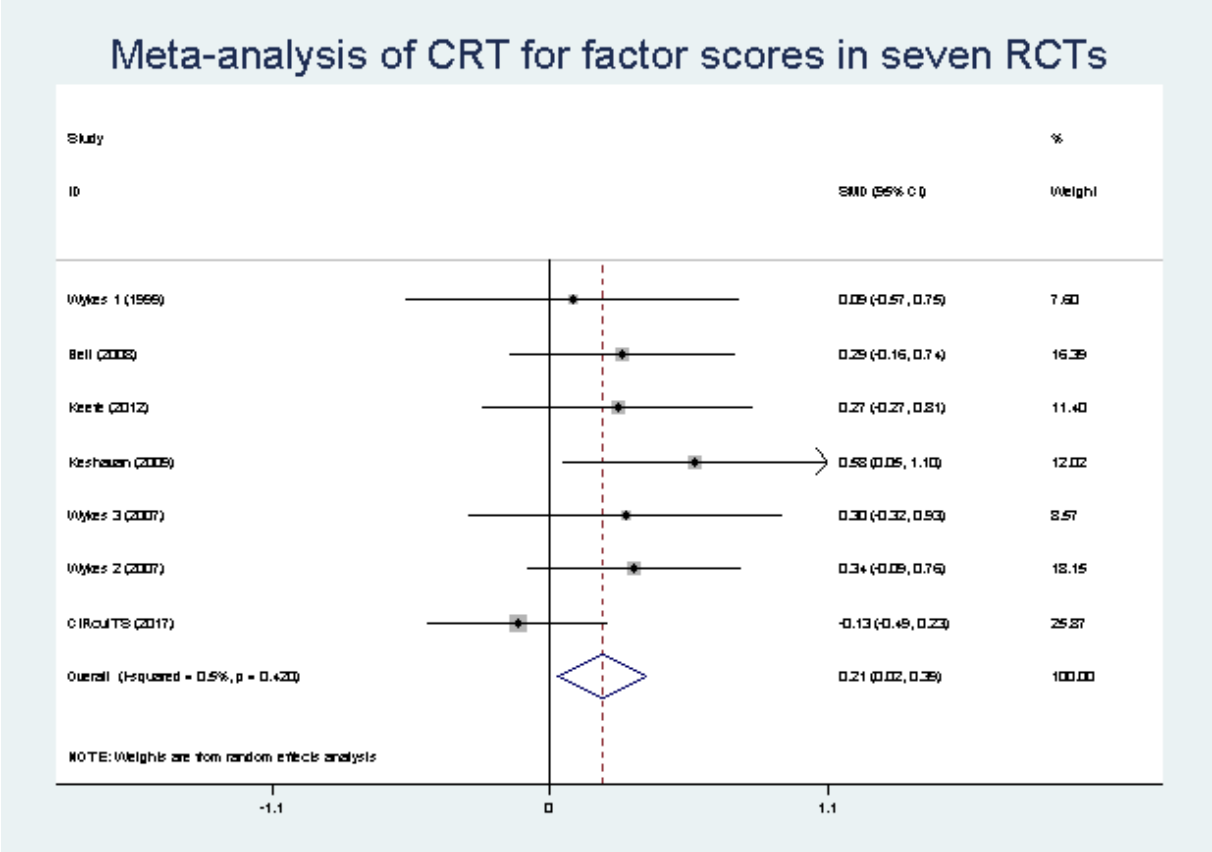



Figure 3.2: LFA model for the latent summary measure of cognitive abilities across 3 time points (six items). Single-headed arrows indicate the hypothesized relationships between items and factor. Standardized loadings are shown next to each path. Double-headed arrows represent covariation between two variables. The correlation coefficient is shown above the double headed arrows. The errors of the item regressions on the factors are omitted. Circles represent latent factors and the rectangles represent the observed items.*= $< 0.001$.

**Factor scores and their meta-analysis: results**

Figure 3.3: Random effects meta-analysis of factor scores




In the next step, I estimated the factor scores for the latent measure of memory, processing speed and executive function separately at baseline and end-of-treatment. The cross-sectional CFAs used to compute the factor scores (FS) at baseline and at the end-of-treatment, on the 6-items structure (LNS, CATFLU, FAS, log(TMTA), log(TMTB) and WCST PE) confirmed in the LFA, had all standardised loadings above 0.4 and showed acceptable fit:

- CFA at baseline (454 observations): RMSEA = 0.064 (90% CI 0.036-0.094), CFI = 0.957 and SRMR = 0.050
- CFA at end-of-treatment (411 observations): RMSEA = 0.058 (90% CI 0.026-0.090), CFI = 0.966 and SRMR = 0.061

The obtained FS were strongly correlated across time ($\rho$ = 0.836) and they were slightly skewed. FS were standardised to mean 0 and variance 1. Looking at the standardised FS means within treatment groups (CRT vs TAU), I can see that cases seem to improve between baseline and end-of-treatment while controls seem to worsen (see Table 3.11).

Table 3.11: Standardised factor scores statistics within treatment group

| Time point | Factor scores | | | |
|---|---|---|---|---|
| | Mean (Var) | | Range | |
| | CRT (n=209) | Control (n=202) | CRT (n=209) | Control (n=202) |
| Baseline | 0.021 (0.988) | -0.022 (1.014) | -3.281 to 2.402 | -4.049 to 2.314 |
| end-of-treatment | 0.099 (0.991) | -0.103 (1.001) | -3.337 to 2.452 | -4.592 to 2.189 |

**Meta-analysis of factor scores: results** Figure 3.3 illustrates a forest plot for cognitive factor scores for memory, processing speed and executive function, depicting an overall significant positive effect of CRT of 0.207 (effect size (ES), 95% confidence interval: 0.024 to 0.390, p-value = 0.027). The $I^2$ was 0.5%, suggesting that almost all variance of the effect size could be explained by sampling variance and not by study differences, which were very small. Also the Q statistic for the test of heterogeneity yielded a p-value of 0.420, confirming that there was no evidence of heterogeneity between trials. This result provides justification to the decision of not including the study site as a covariate in the prediction models I developed.

**Factor scores for Fiszdon (Model 1)** In this paragraph, factor scores for the external validation of the prediction models to develop were computed. Because LNS and CATFLU were not measured in this study and they were needed to compute the baseline and end-of-treatment outcome factor scores (see Section 3.3.1), I used the values for LNS and CATFLU measured on 23 out of the 75 patients before and after receiving CRT in a separate RCT by the same principal investigator Fiszdon ('Fiszdon 2', 'Efficacy of Social Cognition Training in Schizophrenia (DCTRS)', see Subsection 3.1.1).

The mean percentage of missing data at baseline for these six outcomes was 23% (range 0-69), and similarly at the end-of-treatment it was 24% (0-70).

In order to compute the factor scores for Fiszdon's data, Muthén and Muthén on the Mplus discussion suggested to rerun the analysis by inputting the new data constraining all parameters to be fixed as in the 6-item CFA models at baseline and end-of-treatment (B. Muthén and L. Muthén n.d.). The obtained CFAs fits were almost good at baseline, but poor at the end-of-treatment:

- CFA at baseline (75 observations): RMSEA = 0.057 (90% confidence interval: 0.000-0.132), CFI = 0.943 and TLI = 0.939;
- CFA at end-of treatment (64 observations): RMSEA = 0.127 (90% confidence interval: 0.058-0.194), CFI = 0.789 and TLI = 0.774.

The factor scores for Fiszdon were only slightly skewed and again a pattern of improvement in the cases and worsening in the controls could be observed (see Table 3.12)

Table 3.12: Standardised factor scores statistics within treatment group for the study 'Fiszdon 1'

| Time point | Factor scores for the study 'Fiszdon 1' | | | |
|---|---|---|---|---|
| | Mean (Var) | | Range | |
| | CRT (n=41) | Control (n=23) | CRT (n=41) | Control (n=23) |
| Baseline | 0.1272 (0.9124) | -0.2544 (1.1193) | -1.8186 to 2.2052 | -3.3604 to 1.7102 |
| end-of-treatment | 0.2119 (0.8742) | -0.3778 (1.0413) | -1.8214 to 2.0506 | -3.2973 to 1.6046 |

### 3.3.2 MissForest-Lasso precision medicine models: results

#### Step 1: Prediction problem definition and potential predictors inspection

**Model 1**: I aimed to identify moderators of CRT success through the development of a precision medicine prediction model from individual data of the seven RCTs described in the section 3.1.1, using the method MissForest-Lasso studied in the previous chapter 2. The outcome of interest was the summary cognitive measure of memory, processing speed and executive function in patients accessing CRT compared to controls at the end-of-treatment (see section 3.2.1). Predictors were studies and patients characteristics including demographics, quality of life, symptom data, global cognition, the outcome measured before treatment (all baseline variables) and interactions between all these baseline variables and treatment type. For the literature about moderators of CRT, see Section 1.1.3.

**Models 2a and 2b**: the aims were 1) to develop a precision medicine model for CRT success by using the same data and in the same way as Model 1, but with the observed cognitive measure Wisconsin Card Sorting Test number of Perseverative Errors (WCST PE) instead of a latent outcome as a dependent variable; 2) to study the applied performance of MissForest-Lasso with and without missing data in the dependent variable. WCST PE measures executive function and is one of the most clinically important cognitive measure. In our case it was also the cognitive outcome with least missing data (14%). It was included in the development of factor scores for the latent summary measure, but while the latent factor was predominantly correlated with log(TMTB) (standard loadings $> 0.82$), WCST PE was only moderately correlated with the factor (approximately -0.48). WCST PE was therefore substantially different from the factor as a cognitive measure. However, WCST PE was only measured in six out of seven studies ('Keefe' was excluded). The distribution of WCST PE was slightly skewed ($<$

1.5) at both baseline and end-of-treatment, but showed similar distributions across time and treatment arms. A transformation of the outcome was therefore not necessary. Raw statistics showed an improvement (i.e. less perseverative errors) in WCST PE from baseline to the end-of-treatment, which seems to be greater in cases than in controls (see Table 3.13). Baseline WCST PE correlated 0.63 with WCST PE at the end of treatment.

Table 3.13: WCST PE statistics within treatment group

| Time point | **WCST PE** | | | |
|---|---|---|---|---|
| | Mean (Var) | | Range | |
| | CRT (n=212) | Control (n=198) | CRT (n=212) | Control (n=198) |
| Baseline | 29.94 (404.41) | 30.13 (495.76) | 3 to 96 | 4 to 94 |
| end-of-treatment | 24.52 (420.06) | 27.05 (405.86) | 4 to 95 | 4 to 96 |

Model 2a only included individuals having the outcome (N=356) and Model 2b also included individuals in the six studies with missing outcome (N=410) and required the imputation of missing outcome.

#### Step 2: Optimal coding of predictors

The number of categorical and ordinal variables with small number of observations within a level (1 or 2 people, e.g. in the variable 'additional psychiatric diagnosis') was 49. Thus, levels were collapsed in a reasonable way in order to have at least 10 people per category.

#### Step 3: Model specification

After dropping the baseline variables with more than 70% missing data, the mean percentage of missing data (including missing data by design) was 39% among the baseline predictors. The number of study information variables excluded because they were perfectly multicollinear with at least one variable was 44 (not including the dummy-coded terms for the categorical variables), thus only 24 study information variables were left in the model (see Table 3.1). In agreement with the clinician, I also removed six categorical medicine data variables because the number of levels was greater than seven and they were therefore of little predictive usefulness (for example: the different antipsychotic medicine used by patients).

**Model 1**: Model 1 had the summary cognitive measure as a dependent variable, i.e. the factor scores computed in the factor analysis on six cognitive outcomes CATFLU, FAS, LNS, WCST, log(TMTA), log(TMTB) (see 3.2.1). The model's linear predictor included 112 baseline

variables plus the intercept and all the 2-way interaction terms with the treatment variable. The predictors in the model after the above exclusions were the following:

- baseline outcome factor scores
- baseline cognitive outcomes CATFLU, FAS, LNS, WCST, log(TMTA), log(TMTB)
- treatment : 'Comparison condition' - Reference category (RC), 'Cognitive Remediation condition'
- patients demographics (see Table 3.3):
  - gender: 'male' (RC),'female'
  - baseline age
  - race: 'White' (RC), 'Asian', 'Black', 'Other'
  - education category: 'secondary education' (RC),'primary education or less', 'tertiary/further education'
  - education years
  - primary diagnosis: 'Undifferentiated SCZ' (RC), 'Acute and transient psychotic disorder', 'Disorganized SCZ', 'Paranoid SCZ', 'Psychotic disorder suggestive of SCZ', 'Other', 'Schizophrenia'
  - marital status: 'single/unmarried' (RC), 'married', 'separated/divorced/widowed'
  - time since 1st contact: 'more than 10 years' (RC),'6-10 years','2-5 years','1 year'
  - time since 1st hospitalization: 'more than 10 years' (RC),'6-10 years','2-5 years','1 year'
- baseline global cognition variables (see Table 3.3):
  - Ammons Quick Test (AQT) for the Intelligence Quotient (IQ)
  - WAIS digit-symbol
  - WAIS vocabulary
- medication variables:
  - Chlorpromazine (see Table 3.3)
  - Depot (injectable): 'no' (RC), 'yes'
  - Generation: 'Atypical' (RC), 'Typical'

- Symptom data:
  - the 30 PANSS measures (7 positive, seven negative and 16 general measures, Kay, Fiszbein, and Opler 1987)
  - 4 summary PANSS measures (general, positive, negative, total, Kay, Fiszbein, and Opler 1987)
  - 5 summary PANSS factors (negative, excitement, cognitive, positive and depression, Lindenmayer, Bernstein-Hyman, and Grochowski 1994)
- Quality of life measures:
  - 3 RSE (confirmation, deprecation and total score - alternative scoring)
  - 22 SBS measures
- 24 study information variables, see Table 3.1

Therefore, after transforming the categorical variables into dummies, there were 278 covariates in the model.

As a result of the Factor Analysis in the Subsection 3.2.1, the observations used to train the model were 411, each of which had a factor score estimated (i.e. the dependent variable was complete).

**Models 2a and 2b**: the model specification for Models 2a and 2b were the same as for Model 1, apart from the fact that no baseline observed cognitive outcomes were included in the linear predictor but WCST PE, i.e. the values of the dependent variable at baseline. Overall the model had 106 baseline variables, i.e. 266 covariates and was trained on 356 individuals in Model 2a and on 410 individuals in Model 2b.

**Correlation matrix** The correlations between the variables in the model specification were usually small to moderate with only a few strong correlations (see Figure D.1 in the Appendix).

#### Step 4: Model estimation

**Model 1**: The first part of MissForest-Lasso combined method, consisting of the MissForest imputation algorithm (Stekhoven and Buhlmann 2012), computed random forest models with the default settings (37 predictors sampled for splitting at each node of a tree with continuous dependent variables and 10 for trees with categorical dependent variables).

The second part of the method, i.e. the Lasso model, was fitted to the imputed data with bootstrap tuning (100 bootstrap samples and 40 values for the tuning parameter $\lambda$, range 0.001-2.500).

The estimated Lasso coefficients for Model 1 are shown in Table 3.14. The tuning parameters minimising the MSE in the bootstrap resampling with the four different tolerance levels were: best $\lambda = 0.0453$, one SE $\lambda = 0.0554$, 3% tolerance $\lambda = 0.1012$ and 15% tolerance $\lambda = 0.2761$.

The best $\lambda$ model selected 28 covariates (without counting the intercept, including 9 interactions), the one SE model similarly selected 22 (5 interactions), the 3% tolerance model selected 11 (1 interaction) and the 15% tolerance model selected only the baseline outcome.

Table 3.14: **Model 1** estimated coefficients for the selected variables are shown. The estimates corresponding to four different $\lambda$s were given: $\lambda$ minimising the MSE, the 1 SE $\lambda$ and $\lambda$s giving the MSE within 3% and 15% of the minimum respectively. The word 'other' was used to indicate the union of a categorical variable levels for which the dummy was not selected. The colon ':' indicates an interaction. All the interaction terms are shown below the separating horizontal line. The star sign * means that the estimates were less than $|10^{-4}|$ in absolute value: 0<0.0001*<0.0001 and -0.0001<-0.0001*<0. These coefficients can be regarded as neglectable.

| Covariate | Model 1 | | | |
|---|---|---|---|---|
| | **Best $\lambda$** | **1 SE tol $\lambda$** | **3% tol $\lambda$** | **15% tol $\lambda$** |
| Intercept | 2.5832 | 2.3270 | 1.5138 | -0.0038 |
| Age | -0.0031 | -0.0022 | | |
| Race, Black vs 'other' | -0.0082 | | | |
| Education category, primary or less vs 'other' | -0.1443 | -0.1286 | -0.0270 | |
| Education Years | 0.0067 | 0.0042 | | |
| Time since 1st Contact, '2-5 years' vs 'other' | 0.0444 | 0.0403 | | |
| PANSS uncooperativeness (G8) | 0.0062 | | | |
| SBS Slowness | 0.0041 | | | |
| WAIS digit-symbol | -0.0020 | -0.0010 | | |
| log(TMTA) | -0.4719 | -0.4230 | -0.2659 | |
| log(TMTB) | -0.3125 | -0.2813 | -0.1829 | |
| Cognition | 0.3060 | 0.3538 | 0.4957 | 0.5827 |
| Strategy training technique rank order, 3$^{rd}$ vs 'other' | -0.0595 | -0.0385 | | |
| Errorless learning technique rank order, No central vs 'other' | 0.0247 | 0.0399 | 0.0255 | |
| Verbal Memory target rank order, No priority target vs 'other' | 0.0011 | 0.0011 | 0.0001* | |
| 3$^{rd}$ vs 'other' | -0.0014 | -0.0013 | -0.0001* | |
| CR sessions delivered one-on-one, yes vs no | | -0.0001* | | |
| Target follow-up, 24 weeks vs 'other' | -0.0709 | -0.0702 | -0.0384 | |
| no follow-up vs 'other' | 0.0103 | 0.0065 | 0.0001* | |
| CATFLU | 0.0161 | 0.0127 | 0.0014 | |
| LNS | 0.0536 | 0.0484 | 0.0358 | |
| CRT:Race, Other vs 'White or Asian or Black' | -0.0759 | -0.0020 | | |
| CRT:Time since 1st Contact, '6-10 years' vs 'other' | -0.1195 | -0.0572 | | |
| '1 year' vs 'other' | 0.0245 | | | |
| CRT:PANSS social avoidance (G16) | 0.0280 | 0.0194 | | |
| CRT:PANSS suspiciousness (P6) | 0.0056 | 0.0026 | | |
| CRT:SBS laughing | 0.0009 | | | |
| CRT:Metacognitive training technique rank order, No central vs 'other' | 0.0007 | | | |
| CRT:Doctoral-level clinicians, yes vs no | 0.0008 | | | |
| CRT:Duration of CRT (weeks) | 0.0004 | 0.0006 | 0.0002 | |

**Models 2a and 2b**: MissForest was run using the same default settings as in Model 1, only the numbers of predictors sampled for splitting at each node of a tree with continuous dependent variables changed from 37 to 35 because there were less predictors compared to Model 1 (see formula in Subsection 3.2). Also the Lasso had the same bootstrap tuning settings as in Model 1, only the range for $\lambda$ changed: 0.316-5.012.

The estimated Lasso coefficients for Models 2a and 2b are shown in Table 3.16. The best $\lambda$ models selected 12 and 16 covariates (including one interaction, without counting the intercept) for Models 2a and 2b respectively (11 covariates were in common, not the interaction); the one SE $\lambda$ models selected 7 and 8 (no interactions, 6 covariates in common), the 3% tolerance models retained 4 and 5 variables (no interactions, 4 covariates in common) and the both the 15% tolerance models chose only the baseline outcome. The tuning parameters minimising the MSE are presented in Table 3.15.

Table 3.15: Models 2a and 2b tuning parameters.

| Model | Tuning parameter | | | |
|---|---|---|---|---|
| | best $\lambda$ | 1 SE $\lambda$ | 3% $\lambda$ | 15% $\lambda$ |
| **Model 2a** | 1.7317 | 2.4678 | 3.5169 | 5.0119 |
| **Model 2b** | 1.5029 | 2.1418 | 3.0522 | 5.0119 |

Table 3.16: Models 2a and 2b estimated unstandardised coefficients for the selected variables are shown. The estimates corresponding to four different $\lambda$s were given: $\lambda$ minimising the MSE, the 1 SE $\lambda$ and $\lambda$s giving the MSE within 3% and 15% of the minimum respectively. The word 'other' was used to indicate the union of a categorical variable levels for which the dummy was not selected. The colon ':' indicates an interaction. The star sign * means that the estimates were less than $|10^{-4}|$ in absolute value: 0<0.0001*<0.0001 and -0.0001<-0.0001*<0. These coefficients can be regarded as neglectable.The highlighted rows correspond to the covariates selected by both the models.

| Outcome = WCST PE | Model 2a | | | | Model 2b | | | |
|---|---|---|---|---|---|---|---|---|
| **Covariate** | **Best $\lambda$** | **1 SE tol $\lambda$** | **3% tol $\lambda$** | **15% tol $\lambda$** | **Best $\lambda$** | **1 SE tol $\lambda$** | **3% tol $\lambda$** | **15% tol $\lambda$** |
| Intercept | 8.5566 | 13.3485 | 15.2644 | 15.0965 | 10.1934 | 14.2627 | 17.0385 | 15.9746 |
| Age | 0.0240 | | | | 0.0271 | | | |
| Education category, primary or less vs 'other' | 0.5729 | | | | 1.1970 | 0.4715 | | |
| Education Years | -0.0363 | | | | -0.0008 | | | |
| Time since 1st Contact, '2-5 years' vs 'other' | -0.4113 | | | | | | | |
| AQT for IQ | -0.0236 | -0.0427 | -0.0272 | | -0.0255 | -0.0398 | -0.0415 | |
| PANSS abstract thinking (N5) | 1.3109 | 0.7427 | 0.0686 | | 1.2594 | 0.8163 | 0.1455 | |
| PANSS lack of judgement (G12) | | | | | 0.0058 | | | |
| PANSS poor rapport (N3) | 1.1483 | 0.4413 | | | 0.7675 | 0.2976 | | |
| PANSS social avoidance (G16) | -0.1498 | | | | -0.3707 | | | |
| PANSS suspiciousness (P6) | | | | | -0.0600 | | | |
| PANSS lack of spontaneity (N6) | | | | | 0.3273 | 0.0901 | | |
| SBS concentration | | | | | 0.0544 | | | |
| Metacognitive training, No central vs 'other' | -3.2742 | -0.6417 | | | -0.8744 | | | |
| Duration of CRT (weeks) | | -0.0127 | | | -0.0469 | -0.0394 | -0.0098 | |
| Target follow-up, 24 weeks vs 'other' | 2.8428 | 2.6128 | 1.3634 | | 3.7705 | 3.4089 | 2.3948 | |
| WCST Perseverative Errors | 0.4587 | 0.4487 | 0.4282 | 0.3703 | 0.4190 | 0.4141 | 0.4018 | 0.3369 |
| CRT:Gender, female vs male | | | | | 0.4552 | | | |
| CRT:Time since 1st Contact, '2-5 years' vs 'other' | -0.1768 | | | | | | | |

**Step 5: Model performance**

**Data imputation performance** **Model 1**: The performance of MissForest imputation was acceptable: the NRMSE was 0.5172 and the PFC was 0.0529 (see Method's Subsection 3.2.2, Step 6, for benchmarks for good performance). **Models 2a and 2b**: the NRMSEs were 0.4804 and 0.4863, and the PFCs 0.0418 and 0.0378 respectively.

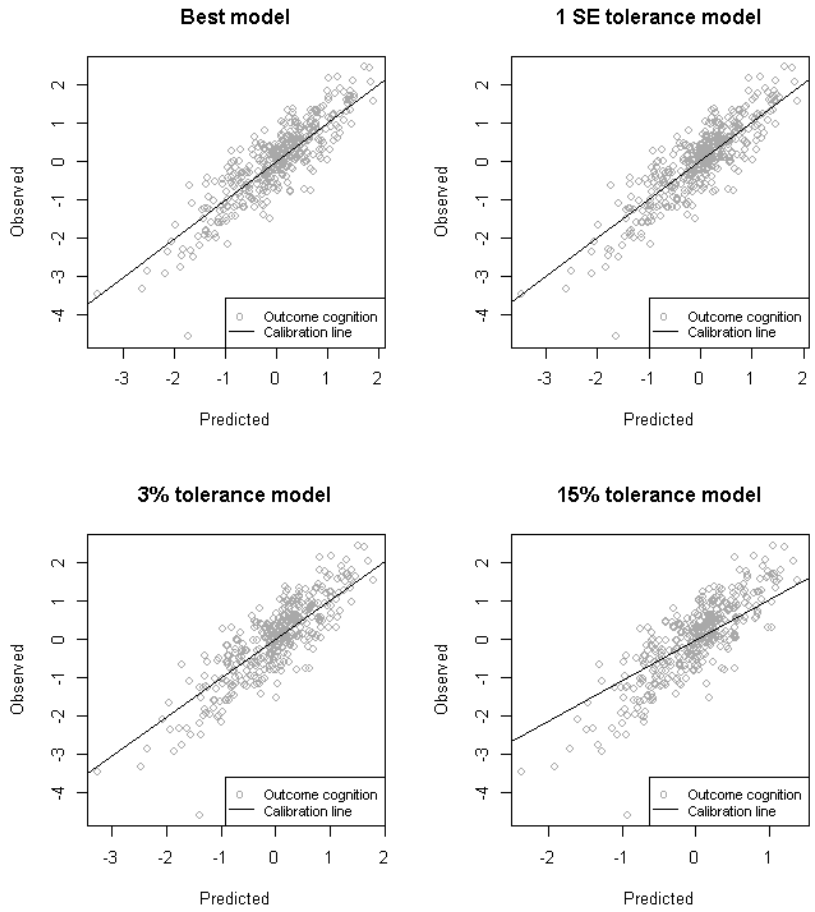
**Apparent performance** **Model 1**: The apparent discriminative and calibrative performance (see Section 1.2.1 for definitions) are shown in Table 3.17. The increase in the apparent MSE with increasing tolerance level was small for the tolerance levels up to 3% (outcome variance=1). In fact, the apparent pseudo-$R^2$ showing a good apparent discrimination for these penalties. Also, the apparent calibration was good: the slope $\beta$ was always just above 1 for the same low tolerance models and the calibration-in-the-large $\alpha$ was close to 0 in all models.

According to the apparent calibration performance, predicted and observed cognition values agreed reasonably well with each other (see Figure 3.4).

Table 3.17: Model 1 apparent performance: the mean squared error (MSE), the pseudo-$R^2$, the calibration slope $\beta$ and the calibration-in-the-large $\alpha$ are shown. The star sign * means that the estimates were less than $|10^{-4}|$ in absolute value: 0<0.0001*<0.0001 and -0.0001<-0.0001*<0.

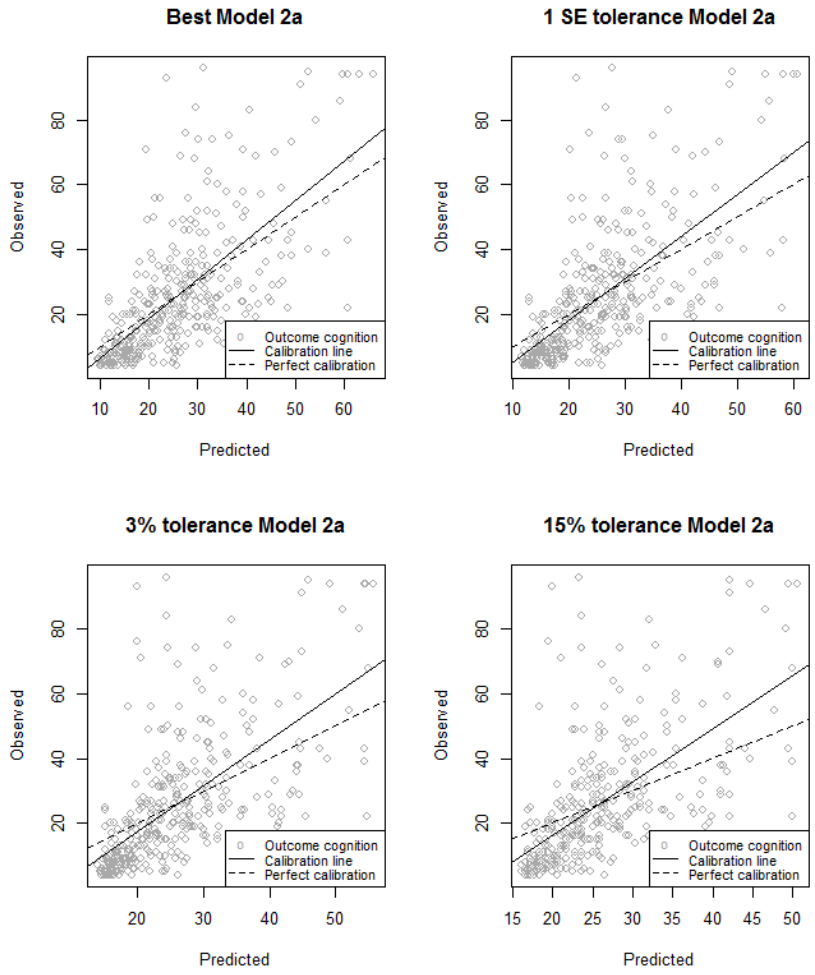
| Apparent Performance | **Model 1** | | | |
|---|---|---|---|---|
| | best $\lambda$ | 1SE $\lambda$ | 3% $\lambda$ | 15% $\lambda$ |
| Apparent MSE | 0.2433 | 0.2539 | 0.2885 | 0.3791 |
| Apparent pseudo-$R^2$ | 0.7567 | 0.7461 | 0.7115 | 0.6209 |
| Apparent $\beta$ | 1.0809 | 1.0929 | 1.1524 | 1.4953 |
| Apparent $\alpha$ | 0.0001* | 0.0001* | 0.0001* | -0.0001* |

Figure 3.4: **Model 1** predictions versus observed outcome values for the best $\lambda$ model, the one SE, the 3% and the 15% tolerance models. Apparent calibration lines are shown.




**Models 2a and 2b**: The apparent discrimination and calibration measures are presented in Table 3.18. The apparent pseudo-$R^2$ indicated that Model 2a explained at most 46% of the outcome variance with the best penalty. The apparent calibration slope ($> 1$) together with the calibration-in-the large $\alpha$ revealed underfitting of the data in all models 2a (see Figure 3.5). The 3% tolerance Model 2a performed slightly worse in prediction accuracy relative to the best Model 2a, compared to the 3% tolerance and best factor scores models 1. The performance was slightly better when the outcome missing data were imputed in Model 2b (see Table 3.18, the apparent calibration lines for Model 2b are in Figure D.2 in the appendices).

Figure 3.5: **Model 2a** predictions versus observed outcome values for the best $\lambda$ model, the one SE, the 3% and the 15% tolerance models. Apparent calibration lines are shown.




| Apparent performance | Model 2a | | | | Model 2b | | | |
|---|---|---|---|---|---|---|---|---|
| | best $\lambda$ | 1SE $\lambda$ | 3% $\lambda$ | 15% $\lambda$ | best $\lambda$ | 1SE $\lambda$ | 3% $\lambda$ | 15% $\lambda$ |
| Apparent MSE | 224.8601 | 238.7501 | 257.3441 | 275.4046 | 199.9324 | 211.7426 | 227.6079 | 253.4311 |
| Apparent pseudo-$R^2$ | 0.4562 | 0.4226 | 0.3777 | 0.3340 | 0.5165 | 0.4879 | 0.4863 | 0.3871 |
| Apparent $\beta$ | 1.2182 | 1.2954 | 1.4085 | 1.6493 | 1.2041 | 1.2759 | 1.3824 | 1.7063 |
| Apparent $\alpha$ | -5.6216 | -7.6117 | -10.5242 | -16.7294 | -5.2209 | -7.0584 | -9.7840 | -18.0706 |

Table 3.18: Model 2a and 2b apparent performance: the mean squared error (MSE), the pseudo-$R^2$, the calibration slope $\beta$ and the calibration-in-the-large $\alpha$ are shown. The star sign * means that the estimates were less than $|10^{-4}|$ in absolute value: 0<0.0001*<0.0001 and -0.0001<-0.0001*<0.

**Step 6: Model validated performance**

**Internal validation** **Model 1**: The bootstrap-corrected estimates of performance are shown in Table 3.19. The pseudo-$R^2$ after correction was only slightly lower than the apparent measure

Table 3.19: **Model 1** bootstrap-corrected performance: the mean squared error (MSE), the pseudo-$R^2$, the calibration slope $\beta$ and the calibration-in-the-large $\alpha$ are shown.

| Performance | **Model 1** | | | |
|---|---|---|---|---|
| | best $\lambda$ | 1SE $\lambda$ | 3% $\lambda$ | 15% $\lambda$ |
| Corrected MSE | 0.2474 | 0.2585 | 0.2929 | 0.3803 |
| Corrected pseudo-$R^2$ | 0.7526 | 0.7415 | 0.7071 | 0.6197 |
| Corrected $\beta$ | 1.088 | 1.0982 | 1.1553 | 1.4959 |
| Corrected $\alpha$ | -0.0017 | -0.0010 | -0.0012 | 0.0002 |

for the best and low tolerance level models, indicating good discrimination. The models also resulted well calibrated after correcting for optimism. Validated discrimination and calibration measures were all only slightly lower than the apparent measures, suggesting that there was little bias due to overfitting.

**Models 2a and 2b**: again there was little estimated optimism so that the bias-corrected performance estimates were very similar to the apparent performance estimates (see Table 3.20). Model performance was better for Model 2b with imputed missing outcomes.

Table 3.20: Models 2a and 2b bootstrap-corrected performance: the mean squared error (MSE), the pseudo-$R^2$, the calibration slope $\beta$ and the calibration-in-the-large $\alpha$ are shown.

| Performance | **Model 2a** | | | | **Model 2b** | | | |
|---|---|---|---|---|---|---|---|---|
| | best $\lambda$ | 1SE $\lambda$ | 3% $\lambda$ | 15% $\lambda$ | best $\lambda$ | 1SE $\lambda$ | 3% $\lambda$ | 15% $\lambda$ |
| Corrected MSE | 225.0474 | 239.3642 | 258.8875 | 275.5891 | 200.2395 | 212.0586 | 227.9758 | 253.4845 |
| Corrected pseudo-$R^2$ | 0.4558 | 0.4211 | 0.3739 | 0.3335 | 0.5158 | 0.4872 | 0.4487 | 0.3870 |
| Corrected $\beta$ | 1.2150 | 1.2938 | 1.4068 | 1.6477 | 1.1990 | 1.2743 | 1.3851 | 1.7133 |
| Corrected $\alpha$ | -5.5555 | -7.5781 | -10.4812 | -16.6936 | -4.9801 | -6.9487 | -9.8222 | -18.2273 |

**External validation** In order to externally validate the models, I used the data from the CRT randomised controlled trial 'Fiszdon 1' (Fiszdon et al. 2016) as test set (75 patients).

Some variables selected by the developed models (see Tables 3.14 and 3.16) were not available in 'Fiszdon 1':

- cognitive outcomes: LNS, CATFLU
- time since 1st contact
- education category

- Ammons Quick Test (AQT) for the IQ
- SBS laughing
- WAIS digit-symbol

I could then only externally validate the 15% tolerance models as they only selected the baseline outcome as predictors (see Table 3.14).

**WCST PE for the study 'Fiszdon 1'** There seemed to be an improvement in the outcome WCST PE at the end of treatment among cases in the study 'Fiszdon 1' (see Table 3.21). The scores in the controls seemed to worsen after TAU and they had very large variance compared to the cases. The correlation of WCST PE measured at baseline with the values measured at the end-of-treatment was 0.8619.

Table 3.21: WCST PE statistics within treatment group for the study 'Fiszdon 1'

| Time point | **WCST PE for the study 'Fiszdon 1'** | | | |
|---|---|---|---|---|
| | Mean (Var) | | Range | |
| | CRT (n=40) | Control (n=23) | CRT (n=40) | Control (n=23) |
| Baseline | 19.5000 (279.6837) | 33.1200 (699.1933) | 4 to 94 | 7 to 94 |
| end-of-treatment | 16.7250 (182.2558) | 35.3913 (736.4308) | 5 to 89 | 5 to 89 |

**Model performance on external data** **Model 1**: With the computed factor scores for 'Fiszdon 1' as outcome and baseline covariate, I ran the 15% tolerance model on the external data and obtained the following measures of performance:

- Pseudo-$R^2$ = 0.6952
- Calibration slope $\beta$ = 1.4847

The baseline outcome alone here explained about 70% of the variance of the outcome. The calibration slope larger than one showed how the model was underfitting the data. This result was not very far from the 15% tolerance original model performance, even though it had better pseudo-$R^2$ (see Table 3.19). However, the internal and external optimisms (see Table 2.1 for definitions) for the MSE, for the calibration slope and for the calibration-in-the-large were very different: the external optimism estimates were much larger than the internal optimism estimates, meaning that applying the model to new data will be likely not to replicate the accuracy results obtained with the original model:

- MSE: internal optimism = -0.0012 vs external optimism = 0.0744;

- Calibration slope: internal optimism = -0.0006 vs external optimism = 0.0106;
- Calibration-in-the-large: internal optimism = -0.0002 vs external optimism = 0.0022.

| Covariate | **Model 1** | |
|---|---|---|
| | Uncalibrated coef. | Re-calibrated coef. |
| Intercept | 1.513 | 1.7477 |
| Education category, primary or less vs 'other' | -0.0578 | -0.0680 |
| log(TMTA) | -0.2659 | -0.3084 |
| log(TMTB) | -0.1829 | -0.2125 |
| Cognition | 0.4957 | 0.5715 |
| Errorless learning technique rank order, No central vs 'other' | 0.0255 | 0.0282 |
| Verbal memory target rank order, No priority target vs 'other' | 0.0001* | -0.0012 |
| Target follow-up, 24 weeks vs 'other' | -0.0384 | -0.0456 |
| no follow-up vs 'other' | 0.0001* | -0.0012 |
| CATFLU | 0.0014 | 0.0004 |
| LNS | 0.0358 | 0.0402 |
| CRT:Duration of treatment | 0.0002 | -0.0010 |

Table 3.22: Final (3% tolerance) **Model 1** uncalibrated and re-calibrated coefficients (coef). The word 'other' was used to indicate the union of a categorical variable levels for which the dummy was not selected. The colon ':' indicates an interaction. The star sign * means that the estimates were less than $|10^{-4}|$ in absolute value: $0<0.0001^{*}<0.0001$ and $-0.0001<-0.0001^{*}<0$.

**Models 2a and 2b**: The 15% tolerance model consisted of intercept and baseline outcome for both Models 2a and 2b (see Table 3.16). The estimated coefficients were slightly different between the two models. The externally validated performances were again poor if I consider the large divergence between internal and external optimism:

- Pseudo-$R^2$ = 0.6096 and 0.5670 for Model 2a and 2b respectively;
- Calibration slope $\beta$ = 2.7004 and 2.9684 for Model 2a and 2b respectively.

### Step 7: Model presentation

**Model 1**: I recalibrated the 3% tolerance model with the optimism-corrected calibration measures in Table 3.19, and presented the 12 recalibrated coefficients in Table 3.22.

The final model variable selection performance suggested that there was one moderator of CRT: the duration of CRT intervention (in weeks). Selected potential predictors of memory, processing speed and executive function were:

- education category ('primary or less' vs 'other')
- 4 out of 6 cognitive outcomes (log(TMTA), log(TMTB), CAFLU and LNS) and the summary measure of the outcomes at baseline (cognition in Table 3.22)

- 3 study information variables (4 covariates):
  - rank order of importance of errorless learning technique ('No central to the intervention' vs 'Central to the intervention')
  - rank of importance for verbal memory target ('No priority target' vs 'Priority target')
  - Target interval between post-treatment and 1st follow-up assessment in weeks (3 levels: 'no follow-up', '24 weeks', '52 and 12 weeks'; selected contrasts: 'no follow-up' vs '52 and 12 weeks', and '24 weeks' vs '52 and 12 weeks')

**Models 2a and 2b**: the 3% models did not select any interaction terms failing to deliver precision medicine predictions (see Table 3.16). The recalibrated coefficients are in Table D.1 in the Appendix. Three of the retained covariates were in common between Models 2a and 2b other then the baseline outcome WCST PE:

- Ammons Quick Test (AQT) for the IQ
- PANSS abstract thinking
- Target interval between post-treatment and 1st follow-up assessment in weeks ('24 weeks' vs '52 weeks, 12 weeks or no follow-up')

Results showed that MissForest-Lasso when imputing missing outcomes slightly outperformed the complete outcome cases MissForest-Lasso in prediction accuracy. Also, variable selection remained stable between the two models. Only Model 2b selected one more potential predictor: the study specific information variable measuring the planned duration of CRT.

### 3.3.3 Secondary analysis: results

The final prognostic (i.e. with assumption of no moderation of treatment) Model 3 (outcome = factor scores), Model 4a (outcome = WCST PE complete) and Model 4b (outcome = WCST PE with missing data) returned similar accuracy and variable selection results to Models 1, 2a and 2b respectively (see Tables 3.23, 3.24 and D.2 and D.3 in the Appendix).

By comparing variable selection and accuracy of Model 3 and Model 1, it was clear that the potential moderator selected by Model 1 (duration of treatment in weeks, see Table 3.22) was very weak. In fact, all variables retained by Model 1 were also chosen by Model 3 apart from the moderator which was not included in Model 3 specification, and the prediction accuracy was very close. All coefficients were similar between models (compare Tables 3.22 and D.2 in the Appendix). Only two more variables (study-specific variables) were selected by Model 3

and not by Model 1: the planned session duration in minutes (90 vs 60) and the indicator for one-to-one delivery of CRT session, although their effect was very weak (see Table D.2 in the Appendix).

The results from Models 4a and 4b confirmed the variable selection of Models 2a and 2b which did not select any interaction term (see Tables D.1 and D.3 in the Appendix)

Table 3.23: **Model 3** bootstrap-corrected performance: the mean squared error (MSE), the pseudo-$R^2$, the calibration slope $\beta$ and the calibration-in-the-large $\alpha$ are shown.

| Performance | Model 3 | | | |
|---|---|---|---|---|
| | best $\lambda$ | 1SE $\lambda$ | 3% $\lambda$ | 15% $\lambda$ |
| Corrected MSE | 0.2542 | 0.2637 | 0.2933 | 0.3803 |
| Corrected pseudo-$R^2$ | 0.7458 | 0.7363 | 0.7067 | 0.6197 |
| Corrected $\beta$ | 1.0840 | 1.0950 | 1.1547 | 1.4959 |
| Corrected $\alpha$ | -0.0017 | -0.0012 | -0.0014 | 0.0002 |

Table 3.24: **Models 2a and 2b** bootstrap-corrected performance: the mean squared error (MSE), the pseudo-$R^2$, the calibration slope $\beta$ and the calibration-in-the-large $\alpha$ are shown.

| Performance | Model 4a | | | | Model 4b | | | |
|---|---|---|---|---|---|---|---|---|
| | best $\lambda$ | 1SE $\lambda$ | 3% $\lambda$ | 15% $\lambda$ | best $\lambda$ | 1SE $\lambda$ | 3% $\lambda$ | 15% $\lambda$ |
| Corrected MSE | 225.0474 | 239.3642 | 258.8875 | 275.5891 | 195.5025 | 209.2188 | 223.5056 | 255.1736 |
| Corrected pseudo-$R^2$ | 0.4558 | 0.4211 | 0.3739 | 0.3335 | 0.5272 | 0.4940 | 0.4595 | 0.3829 |
| Corrected $\beta$ | 1.2150 | 1.2938 | 1.4068 | 1.6477 | 1.1805 | 1.2598 | 1.3731 | 1.7194 |
| Corrected $\alpha$ | -5.5555 | -7.5781 | -10.4812 | -16.6936 | -4.5059 | -6.5711 | -9.5142 | -18.3826 |

External validation for Models 3, 4a and 4b was done on the 15% tolerance models because again some of the model selected variables were not available in the external data from the study 'Fiszdon 1'. As the 15% tolerance models only had the intercept and the baseline outcome in the linear predictor similarly to the Models 1, 2a and 2b, and the estimated coefficients were very similar, the externally validated performance for Models 3, 4a and 4b was very similar (poor) as the one of Models 1, 2a and 2b.

### 3.3.4 Results: summary

The developed precision medicine Model 1 with a summary measure of executive function, processing speed and memory as outcome had good internally validated discrimination and calibration (i.e. the 3% tolerance $\lambda$ model performance in Table 3.19). However, only one potential moderator was selected: duration of treatment in weeks. Moreover, the final Model 1 could not be externally validated because some of the variables in the linear predictor were not

measured in the external data study. External performance for the model corresponding to the 15% tolerance $\lambda$ was too optimistic compared to the internally validated performance.

When the outcome was the observed executive function measure WCST PE, Model 2a (complete outcome) and Model 2b (individuals with missing outcome included) did not show good performance (see 3% tolerance $\lambda$ model performance in Table 3.20), failed to be validated externally and did not select any moderators. However, the variable selection was consistent between these two models (see Table D.1) and the prediction accuracy was slightly better for Model 2b with missing outcome, meaning that MissForest imputation adjusted for some of the bias induced in the complete outcome cases analysis.

Assuming absence of moderation, the prognostic Models 3, 4a and 4b (the outcomes being the summary cognitive measure, WCST PE complete cases and WCST PE with missing data respectively) performed similarly to the corresponding Models 1, 2a and 2b in both prediction accuracy and variable selection. This result suggested that the only selected putative moderator was weak and not of clinical importance.

There was one variable selected in all six models (common among the potential predictors of summary factor scores and WCST PE): the study information variable measuring the target interval between post-treatment and 1st follow-up assessment in weeks.

## 3.4 Discussion and conclusions

The purpose of this chapter was to identify moderators of Cognitive Remediation Therapy (CRT) response in people with SCZ and to develop a robust precision medicine prediction model using individual participant data from seven randomised controlled trials (RCT). The potential predictors and moderators were baseline variables: patients demographics, cognition, symptoms, quality of life, study information characteristics and all their interactions with treatment type. Study information variables were included in order to adjust for any potential between study heterogeneity.

Three different MissForest-Lasso precision medicine models were developed: Model 1, Model 2a and Model 2b. The initial model specifications were the same, but Model 1 had a summary measure of executive function, processing speed and memory as outcome, while Model 2a and Model 2b had the executive function measure Wisconsin Card Sorting test Perseverative Errors (WCST PE) as a dependent variable. Model 2a performed a complete outcome analysis and Model 2b also included individuals with missing outcome in order to compare MissForest-Lasso performance with and without missing outcome values. Only Model 1 in-

cluded a moderator, while the other two models were purely prognostic. While Model 1 showed good internally validated prediction accuracy, Models 2a and 2b performed poorly. None of the models were successfully validated externally due to limited data.

The factor scores of the summarised measure employed as outcome in Model 1 were computed using separate cross-sectional confirmatory factor analyses (CFAs) at each of the 3 different time points. The CFAs had a fixed factorial structure across time which was previously tested for longitudinal invariance. The 1-factor CFAs presented very good fits, but reliability and validity were not tested (see Subsection 3.2.1). The meta-analysis of factor scores presented a significant effect size result (0.21, 95% confidence interval: 0.02 to 0.39), while no one of the six cognitive outcomes implicated in the computation of the factor scores showed any significant effect of CRT at the end-of-treatment in the seven single study published analysis results (Wykes, Reeder, Corner, et al. 1999, Wykes, Reeder, Landau, Everitt, et al. 2007, Wykes, Newton, et al. 2007, Bell et al. 2008, Keefe et al. 2012, Keshavan et al. 2008 and Reeder et al. 2017). The latent variable estimated factor scores summarising these six outcomes seemed to be more precise and capture the positive overall effect of CRT in the different domains (memory, processing speed and executive function) considered altogether. This significant positive effect of CRT for the factor scores in the seven trials meta-analysis was consistent with the 40-studies meta-analysis by Wykes, Huddy, et al. (2011), which showed a significant effect size of 0.45 (95% confidence interval: 0.31–0.59) for the outcome global cognition. However, there was no evidence of heterogeneity of outcome between the seven studies in the meta-analysis of factor scores, as in contrast there was in the meta-analysis by Wykes, Huddy, et al. (2011). In fact, the factor scores prediction model only identified one weak moderator.

The only potential moderator of CRT selected by Model 1 was the study information variable measuring the planned therapy duration. This variable was considered as a putative treatment moderator also in the literature (McGurk et al. 2007 and Wykes, Huddy, et al. 2011). The recalibrated model (see Table 3.22) indicated that planned shorter duration of CRT seemed to be associated with increased cognitive abilities at the end of treatment. However, by rerunning Model 1 without interaction terms in the model specification (no moderation assumption, see Model 3 in Subsection 3.3.3), the prediction accuracy and the variable selection results were very similar. This showed the weakness of the moderation effect. Since planned therapy duration can be seen as a proxy for study quality and had a small effect, it was of little use for a precision medicine model.

Some of the selected potential predictors in Model 1 were widely acknowledged by previous research, in particular:

- higher baseline cognitive status is predictive of improved cognition at the end-of-treatment (Kurtz et al. 2009, Lindenmayer, Ozog, et al. 2017). Not only were the baseline factor scores selected, but also four out of six of the single observed outcomes from which the factor scores were computed were potential predictors: log(TMTA), log(TMTB), CATFLU and LNS.

- education category contribution in the prediction suggests that those who have higher levels of education may be more responsive to treatment (Lindenmayer, Ozog, et al. 2017, Barnett et al. 2006, Koenen et al. 2009 and Ramsay et al. 2018)

Model 1 also selected three study characteristics variables as potential predictors:

- Rank order of verbal memory target (no priority vs priority target): when verbal memory was one of the least important treatment targets, the patients from that study seemed to have lower end-of-treatment cognitive improvement compared to studies having verbal memory as priority target as most important technique in the intervention. From a clinical point of view, this may indicate that targeting memory in the treatment delivery has a good repercussion on the recovery of cognitive abilities. However, only the study 'Bell' had verbal memory as the least important treatment target. When characteristics are unique to one study, then it is not possible to distinguish between study site and study characteristics if there is no variation within study.

- Rank order of errorless learning technique: if the errorless training technique was not central to the intervention in a trial, CRT seemed to have better results on cognition than in the trials where such a technique was the most important in the treatment delivery.

- Target interval between post-treatment and 1st follow-up assessment in weeks (3 levels: 'no follow-up', '24 weeks', '52 and 12 weeks'): the two selected contrasts (see Table 3.22) suggested that trials having planned a follow-up after 12 or 52 weeks, delivered a more effective treatment on the cognitive outcome compared to trials with no follow-up or with 24 weeks between end-of-treatment and follow-up. This variable was selected in all models and could be seen as a proxy for methodological trial quality.

The study characteristics variables seemed to adjust well for differences between trails.

It is interesting to note that the variable selection performance of all models suggested a weak effect of CRT on the outcome measure as the treatment type variable was never selected. This is consistent with the small significant effect size of CRT effect in the seven studies' meta-analysis.

According to the simulation study in Chapter 2, the developed MissForest-Lasso models having a model specification of 278 covariates for Model 1 (411 observations) and 266 covariates for Model 2a/b (356 and 410 observations respectively), with a correlation matrix of mainly low correlations between variables (see Figure D.1 in the Appendix) and a mean percentage of missing data of approximately 40%, should have an almost acceptable variable selection performance and a poor prediction accuracy. The tendency for these models would be underfitting the data by delivering more parsimonious models in terms of selected variables. However, being 3% tolerance models, these models have been shown to yield a positive predictive value of selection of approximately 80%. Therefore, I assume that the selected variables are predictors of the outcome with 80% probability. On the other hand, the underfitting nature of Models 2a/b (Model 1 resulted in good prediction accuracy) should promote generalisability to new data. This might explain why the external validation of the developed 15% tolerance models was optimistic compared to the internal validation.

The Lasso has recently been used to examine predictors of cognitive improvement in response to CRT for patients with recent onset SCZ in order to overcome the problem of overfitting a model by including a large number of predictors relative to sample size (Ramsay et al. 2018). However, the analysis was run on a single study data of 42 patients all undertaking the active treatment, with only 10 pre-selected covariates in the model. Therefore, the analysis had less power than the present analysis and could not identify potential moderators. Moreover, Ramsay et al. 2018 worked on complete datasets without tackling the problem of missing data as I did.

Limitations for this analysis were the following:

- I could not assess measurement invariance of the latent summary measure of cognition factorial structure also across studies as I did across time, because of limited data (large proportion of missingness by design). This analysis would have given a more robust structure for our latent outcome. Also the CFAs were not conducted on independent datasets, as it should be done to achieve more validity.

- Models 2a (only individuals with the outcome) and 2b (also individuals with missing outcome) had similar variable selection performances and the prediction accuracy of Model 2b was slightly better than Model 2a, suggesting that MissForest-Lasso corrected for the complete outcome cases analysis bias. However, Chen and Wang (2013) in their simulation study showed that deleting incomplete observations in a complete cases Lasso analysis yielded not only a much larger prediction error than MI-Lasso (group Lasso penalty

combined with MICE), but also lower sensitivity of selection, especially when missing data were MAR. In our case, Model 2a was only restricted to complete outcome cases and not to all complete cases, so that the bias induced in the analysis was lower compared to a complete cases analysis. A simulation study comparing complete outcome cases analysis with outcome imputation analysis for MissForest-Lasso would be needed to formally assess variable selection in this case.

- The large amount of missing data due to drop out and by design certainly introduced bias in all the analyses run in the chapter (FA and MissForest-Lasso model). In the FA, the factor scores were estimated with the Bartlett Method, which is based on regression estimators and is not very efficient in presence of missing data. Factor scores estimated with expected posterior-weighted (full) maximum likelihood methods would have been significantly more reliable than regression estimators when large percentages of missing data were present (Estabrook and Neale 2013). However, this technique was not implemented in the softwares available for the project. MissForest-Lasso showed relatively good performance if data were MAR, but I cannot exclude MNAR data, as missingness due to drop out will induce more bias under this assumption. A further simulation study scenario analysing MissForest-Lasso under the MNAR assumption would be useful to understand how biases are managed.

- There was some bias in the estimation of the factor scores using the Bartlett method, because I developed the cross-sectional CFA models using the factorial structure suggested by the longitudinal factor analysis, in which baseline informed the factor model at end-of-treatment and follow-up. As a results, internal validated estimates of prediction accuracy for the develop model might still be too optimistic.

- As factor scores for the outcome could have been computed also at follow-up, a longitudinal MissForest-Lasso prediction model could have been developed. The advantage of such a model would be the analysis of the CRT durable effect on cognition. However, only experimental R packages running longitudinal regularised regression were available at the time of the analysis. Therefore, I limited my model to only predict end-of-treatment outcome from baseline scores. Recently, the promising R package `lmmen` has been published (2017) and it will be possible to develop longitudinal lasso models to predict CRT cognitive improvement at each time point longitudinally in order to emphasize the durability of the CRT effect on cognitive abilities (McGurk et al. 2007, Wykes, Huddy, et al. 2011, Fiszdon et al. 2016)

- Because of lack of data, a proper leave-site-out validation (Steyerberg and Harrell 2016) was not run. It is recommended that a leave-site-out validation is conducted to explore geographic transportability of the model results.

- A rigorous clinical usefulness analysis for continuous outcomes need to be developed in order to understand whether the developed model is more useful than the 'treat all' or 'treat none' default policies. However, due to the poor prediction accuracy of Models 2a/b and the failed external validity for Model 1, a clinical usefulness analysis would not be meaningful.

- A limitation for assessing variables for clinical practice was not being able to provide inference estimates for Lasso. Although inferential statistics for Lasso have been proposed (Lockhart et al. 2014 and Hastie 2015), this inference estimates are vulnerable to substantial bias and not yet safe to use from the current experimental statistical packages (e.g. R package *covTest*). Therefore, there is not evidence for the model selected variables to be significant predictors or moderators of CRT. They can only be treated as potential predictors and moderators to be tested in future research. Recently, permutation tests p-values have been developed which can be used for such purposes (Arbet et al. 2017).

To the best of my knowledge, there is no literature about precision medicine models for psychiatric data able to identify moderators of treatment, even though moderators were found in meta-regression analyses (McGurk et al. 2007 and Wykes, Huddy, et al. 2011, see Subsection 1.1.3). For precision medicine prediction to work, it is recommended that more data modalities are used in the model specification, such as brain imaging data, genetics, OMICS data altogether with demographics, symptoms, medicine and quality of life data (Bzdok and Meyer-Lindenberg 2018 and Eyre, Singh, and Reynolds 2016).

### 3.4.1 Conclusion

It is safe to conclude that the developed MissForest-Lasso prediction medicine model (Model 1) had good internally validated discrimination and calibration. However, there was not enough signal in the data for moderation, and external validation and impact assessments are still required before the model is deployed for use as a decision support tool. The strongest prognostic factor was the baseline outcome.

Future studies should use a reliable and validated scale of cognitive outcomes and include putative predictors and moderators from a larger range of data modalities.

# Chapter 4

# Final discussion and conclusion

There has been considerable progress in clinical psychology and psychiatry in recent decades (Dwyer, Falkai, and Koutsouleris 2018). A large number of risk prediction models using statistics and machine learning have been proposed with good internal and external validated performances (Bernardini et al. 2017). For example, criteria have been developed and validated for prevention of schizophrenia (SCZ), to identify individuals at risk of onset of psychosis (i.e. clinical high-risk or prodromal adolescents and young adults) and to follow them over time (Koutsouleris et al. 2016). However, psychotherapeutic or pharmaceutical treatments for mental illnesses are generally effective in only 30–50% of patients (Dwyer, Falkai, and Koutsouleris 2018). Therefore, a shift towards tailoring psychiatric treatment for individual patients or subgroups of patients with similar characteristics is needed through precision medicine (Wium-Andersen et al. 2016). Nevertheless, to date only a few precision medicine models are known in psychiatry and, to my knowledge, none have been developed for SCZ's psychological treatment Cognitive Remediation Therapy (CRT, Wykes, Brammer, et al. 2002). Precision medicine models would predict CRT heterogeneity among patients and allow a patient to be assigned the most likely best treatment.

This PhD project aimed to improve the methodology for precision medicine models and to apply them to a clinical data set. Specifically, the two main purposes of the project were:

- to develop a precision medicine prediction model using statistical learning methods (Hastie, Tibshirani, and Friedman 2008) combined with imputation techniques able to yield good variable selection performance and to deal with large percentages of missing data in the predictors, and lower percentages of missing data in the outcome,
- to find moderators of CRT in people with SCZ using individual participant data of multiple randomised controlled trials.

In attempting the above aims, I had to address the following statistical problems: having a large number of variables relative to sample size, overfitting, multicollinearity, variable selection or measurement of variable importance in the model, dimension reduction of commensurate outcomes and longitudinal invariance of a latent factor.

Using simulations mimicking a variety of plausible settings of data in clinical trials (see Chapter 2), I compared the accuracy and variable selection performance of different modelling techniques combining statistical learning methods such as Least Absolute Shrinkage and Selection Operator (Lasso, Tibshirani 1996), Elasticnet (Zou and Hastie 2005), Random Forests (Breiman 2001) and Conditional Inference Random Forests (Strobl, Boulesteix, et al. 2008) with two missing data imputation techniques: Multivariate Imputation using Chained Equations (MICE, Van Buuren and Oudshoorn 2000) and the non-parametric Random Forests imputation MissForest (Stekhoven and Buhlmann 2012). I demonstrated that MissForest imputation was superior in handling missing data than MICE. The combined methods MissForest-Lasso and MissForest-Conditional Random Forests yielded the best prediction accuracy performance among the other methods when in presence of high percentages of missing data in the predictors (up to 40% on average) and lower percentages of missing data in the outcome (up to 20%), especially when variables were strongly correlated. These two methods outperformed Lasso and Elasticnet combined with MICE, and Elasticnet combined with MissForest. MissForest-Lasso and MissForest-Conditional Random Forests performance demonstrated a good trade-off between prediction accuracy and interpretability as the Lasso naturally performs variable selection and Random Forests returns a measure of variable importance according to the feature contribution in predictions. MissForest-Conditional Random Forests always ranked the true predictors as top variables.

Hastie et al. (2008) suggest the use of the one-standard error penalty (also called tolerance penalty) for Lasso regression models as a rule of thumb to achieve better variable selection, instead of the use of the best penalty minimising the error. However, they did not provide any mathematical proof or experimental justification for the rule. Musoro et al. 2014 also advised to choose the Lasso model corresponding to an even stronger penalty, giving an error within 3% of the minimum, in order to correct for model selection inconsistency. They presented a simulation study to support their recommendation. In my simulations, I compared both one-standard error and 3% tolerance penalties variable performance with the best penalty and a 15% tolerance penalty. The simulation results suggest that when the vector of true predictors is sparse and there are many variables relative to sample size, the so called 3% tolerance penalty model performs better in variable selection than the one-standard error penalty recommended

by Hastie et al. (2008) with similar prediction accuracy.

Both Lasso and Random Forests (like most statistical learning methods) allow the analyses of a large number of variables relative to sample size, therefore these methods could be used to develop precision medicine models for CRT, including all the psychiatric baseline variables (demographics, symptoms, medicine, quality of life, study information, baseline cognitive outcomes) and their interactions with treatment type even though the sample size was relatively small (278 covariates vs 411 observations). Bias due to overfitting in the models was corrected with bootstrap internal validation as for Harrell, Lee and Mark (1996). However, if there was strong multicollinearity (correlation of 0.8 between variables), the Lasso's variable selection performance decreased and Conditional Random Forests worked better in ranking all true predictors as most important variables. From my simulation results, I deduce that an ideal correlation matrix for a reliable feature selection for Lasso is given by low to moderate correlations among true predictors and between true predictors and noise variables.

In the simulations with strong correlation (0.8) between predictors and irrelevant covariates, MissForest-Lasso feature selection performed inconsistently, but with good prediction accuracy. This happened because, if noise variables are highly correlated with the true predictors, their correlations with the outcome are on average almost as strong as the true predictors' correlation with the outcome. Random sampling may cause that false predictors are slightly better ones and then tend to be selected by Lasso model selection. As the wrongly retained fake predictors are still predictive (even in a new data set), a good prediction accuracy is still obtained, which is similar to MissForest-Conditional Random Forests accuracy.

MissForest-Lasso was chosen as the most suitable method to develop a precision medicine model for CRT as, when the number of potential predictors is large, the variable selection process allows clinical use and interpretability. For example, when the analysis involves data obtained through questionnaires and multiple time consuming individual tests, MissForest-Lasso variable selection will allow the number of tests to be measured on new patients by clinicians for predicting new outcomes to be reduced and more manageable. In contrast, MissForest-Conditional Random Forest requires all the variables in the initial model specification to be measured in order to make predictions and be of use for clinicians. Therefore, it is not an ideal method for prediction, as it is often not feasible or too expensive collecting large amounts of data and patients can only be assessed for a limited amount of time. MissForest-Conditional Random Forests would be more suitable for OMICS and MRI data analysis. However, it is important to notice that, in this PhD project, these methods were not assessed for extremely large datasets as bioinformatics, neuroimaging or genetics data, where thousands of variables with

relatively small sample sizes are involved.

I identified an error in the code of MICE-Lasso by Musoro et al. (2014, see the Appendix Section B.1) and the replication of their simulations showed that MICE-Lasso selection performance was not as good as published. The failed variable selection performance of MICE-Lasso in presence of missing data was due to averaging the selected coefficients across imputed datasets, which resulted in a final model including almost all variables, without distinction between true predictors and noise variables. This happened because the false predictors with missing data were made more correlated to the outcome through MICE and were therefore selected. A secondary analysis running multiple MissForest-Lasso imputation was performed to compare variable selection with MICE-Lasso for some scenarios (see Figure A.4 comparing single MissForest imputation with ten MissForest imputations in the Appendix). The induced correlation between variables with missingness and the outcome was reduced in MissForest imputation compared to MICE, as discussed in Section 2.3. This confirms the superiority of the MissForest algorithm, which predicts missing data in a more accurate way than MICE, through iterative random forests models, when data are complex. For example, Shah et al. (2014) suggest the use of MissForest instead of MICE for complex epidemiological data.

I applied MissForest-Lasso to develop a precision medicine model for CRT using individual participant data from seven RCTs. I successfully conducted a factor analysis to obtain factor scores from a latent summary measure of commensurate cognitive outcomes as a dependent variable. Although reliability was not measured for the factorial structure of the latent measure, longitudinal invariance was tested and the structure was confirmed across time. My results showed that the 3% tolerance MissForest-Lasso model had good internally validated performance and selected only one potential moderator of treatment: the planned duration of CRT. This variable was considered as a putative moderator in the literature (McGurk et al. 2007 and Wykes, Huddy, et al. 2011). However, secondary analyses suggested that this potential moderator effect was very weak and that the model could not be considered a precision medicine model.

## 4.1 Limitations

### 4.1.1 Simulation study drawbacks

In the simulation methods, the outcome (complete or with missing data) was always used to impute the missing data in the covariates through both MICE and MissForest imputation in the model development. However, the inclusion of the outcome in the imputation model to

develop a prediction model can be problematic. If the aim is predicting the outcome given the covariates, filling in missing data in the covariates exploiting their correlation with the outcome can introduce bias (optimism of the performance estimates). For this reason, Kuhn et Johnson (2013) strongly advise not to include the outcome in the imputation model and to use imputation methods (KNN, Troyanskaya et al. 2001 or bagged-trees imputation, Kuhn and K. Johnson 2013) which allow the missing data in the test set to be imputed from the information given by the training set covariates only. The latter imputation methods can only be used when there are enough complete records in the training dataset, which was not in this case. When I internally or externally validated the models in the simulation study, the test or external data were always imputed according to an imputation model iteratively built on themselves and not on the training data. This is because MICE and MissForest specifically impute their input dataset for their complex algorithmic nature. This is also the reason why the accuracy of the imputation for MICE or MissForest is superior to other imputation methods. However, I applied bootstrap internal validation in order to correct for bias in the estimation of the prediction accuracy performance.

In classical statistics it is valid to use multiple imputation for the outcome measure, but depending on the assumptions it can be risky (Van Buuren 2007). When all covariates are complete, and the outcome is incomplete, a correct imputation model will yield valid inferences on the parameter estimates from the imputed data when accounting for the random error (Rubin 1981). Under a missing not at random (MNAR) assumption (i.e. missingness depends on the unobserved data), the inferences obtained from just the complete records may be wrong. Therefore, the imputation of the outcome is useful when we know or suspect that the data are MNAR. Under the missing at random (MAR) assumption (i.e. missingness depends on the observed data), there is no advantage to impute the outcome when the imputation model and the substantive model coincide (i.e. the variables used for imputation are the same as included in the main model), and the results may have large variance because of simulation error when the number of imputations is small (Van Buuren 2007).

In prediction modelling, the NMAR assumption always challenges missing data imputation techniques. I did not consider missing data under an NMAR assumption in the simulation study and recommend to include such scenario in future studies to analyse the consequent prediction accuracy and variable selection of the combined methods. Imputing data assumes MAR and this assumption is often reasonable if a large number of variables are used for the imputation model (Hippel and Lynch 2013). However, the best way to assess prediction bias is external validation. If the combined model predicts well on new cases, the imputation step seems to work.

In the case of MissForest, I adopted single imputation. In classical statistics, single imputation can give correct estimation of model parameters when there is no mean-variance relationship, but standard error estimates are biased towards zero, inflating type I error rates. This is because single imputation is 'optimistic' about the extent of error that would be observed. Multiple imputation for continuous covariates accounts for the random error we would have observed, if we had retroactively measured these missing values. The Expectation Maximisation (EM, Dempster, Laird, and Rubin 1977) algorithm works in a similar way by averaging over a range of possible observed outcomes. Multiple imputation is a process of iteratively generating additive error for conditional mean imputation, so that through a small number of imputations, models and their errors can be combined to get correct estimates of model parameters and their standard errors.

In prediction modelling, bias due to missing data uncertainty can be corrected through optimism-correction via bootstrap internal validation including the imputation step in the resampling scheme (Shao and Sitter 1996). However, the statistical learning imputation method MissForest, by averaging over many unpruned classification or regression trees, intrinsically can be seen as a multiple imputation scheme (Stekhoven and Buhlmann 2012). Nevertheless, if the aim is assessing uncertainty of MissForest imputed values, another simulation study with a multiple MissForest imputation scheme could be performed, where variables are selected via the penalty Group-Lasso in order to manage the different imputed datasets variable selection in a simple and sensible way (Chen and Wang 2013).

Since MICE is an iterative method based on Gibbs sampling, it is important to monitor convergence of chains (sequences of draws of parameters and data from conditional distributions, Van Buuren and Oudshoorn 2000). To assess convergence of multiple chains, the stationariness of each chain by the end of the default burn-in period (5 iterations for the R package `mice`) needs to be examined visually, by looking at trace plots of summaries of the imputed values distribution (means and standard deviations) against iteration numbers. Also, looking at a small number of different chains (imputations) for each imputed variable using a different set of initial values for each imputation is a further check of convergence and stability of the algorithm. However, these checks were not done in our simulation study because of the time required to do so.

As the selection of true moderators was an important aim in our thesis, the lack of information on how MissForest-Random Forest ranked true moderators in the variable importance measure in the 20-covariate scenario is a limitation of the simulation study.

Alternative models worth to be assessed in future simulation studies would be Bayesian

Lasso (Park and Casella 2008) and Bayesian Networks (BNs, N. Friedman, Geiger, and Goldszmidt 1997). Unlike Lasso, Bayesian Lasso provides inference estimates on the selected parameters to guide variable selection, which is performed using both Bayesian and likelihood approaches. BNs is a machine learning algorithm which explicitly models the interdependence between the variables under investigation through Bayesian learning algorithms that discover the Bayesian structure and/or estimate parameter values from data. Like the Lasso and Bayesian Lasso, BNs can perform variable selection (Tawfik and Goodwin 2004). Like Random Forests, BNs can manage missing data (Niloofar and Ganjali 2014). Unlike RF and the Lasso, BNs are able to represent uncertainty and causality. BNs decision models have already been successfully used in psychiatry for supporting the diagnosis of dementia, Alzheimer's disease and mild cognitive impairment using multiple study datasets with different attributes. This analysis overcame problems with studies where some domains were measured with different questionnaires (Seixas et al. 2014).

### 4.1.2 Limitations of precision medicine model development

I developed a precision medicine model with a latent summary measure of executive function, processing speed and memory cognitive outcomes as the dependent variable. Factor scores for this latent factor were computed in an unbiased way through the Bartlett method with respective estimated standard errors. However, I used the factor scores as our outcome without accounting for their standard errors and this generates bias. Factor scores computation was done once for all in the factor analysis prior to developing the prediction model, without making this scores estimation part of the bootstrap resampling process for internal validation, which would have corrected for this bias. For this reason, I made sure that the two cross-sectional confirmatory factor analysis (CFA) models, from which the scores were estimated, had very good fits. Standard errors of factor scores for an individual are larger when there are missing items (B. Muthén and L. Muthén n.d.), thus the high percentage of missingness by design in our data influenced negatively the size of standard errors even though the CFAs fitted the data well.

A correct bootstrap internal validation process for the factor scores would have estimated the factor scores for each bootstrap dataset by generating different outcomes for each bootstrap dataset model. However, this would have been done prior to MissForest imputation and there would have been convergence problems in the factor analysis for a large proportions of bootstrap datasets because of the missing data.

Another solution would have been using recently developed regularised structural equation

models (Jacobucci, Grimm, and McArdle 2016), which would have accounted for all forms of error and the imputation method would have been EM. Nevertheless, the problem with our data remains the high percentage of missing data, thus also a regularised structural equation model could have had convergence problems, which do not happen with MissForest.

In the precision medicine model I did not use a validated and reliable scale developed with robust psychometrics methods for the dependent variable. This was not feasible, as it often takes years and a large research team to develop an acceptable scale (Quirk et al. 2013).

In this thesis, the developed models did not necessarily include the main effects of the interaction terms. This is the case in Lasso or Elasticnet regression when the coefficients are shrunk to zero and are not included in the final model.

In inferential statistics, models with interaction effects should include the main effects of the variables that were used to compute the interaction terms, regardless of the coefficients being significantly different from zero or not. even if these main effects are not significantly different from zero. Otherwise, main effects and interaction effects can get confounded: for example, if one of the main effects is not in the model, the significance of the interaction could be due to the absent main effect and not be reflective of an actual interaction. In fact, arbitrary changes in the zero point of the original variables can result in important changes in the apparent effects of the interaction terms.

However, in prediction modelling, the main purpose is to predict well and not to give a causal interpretation. Lasso selects what is informative for prediction. If we consider Lasso as a method to get the highest predictive performance with the smallest number of features, it is acceptable that Lasso selects an interaction term but not the main effects. It simply means that the main effects are not informative (their effect is zero), but interactions are (Hastie 2015).

Not including the main effects corresponding to the selected interaction terms in the CRT Lasso model is therefore not a problem if the analysis aim is prediction modelling. However, this piece of research was also interested in the interpretability of the model for use in clinical practice, and excluding the main effects from the model limits an interpretation to this purpose. It is possible to either include or exclude both main effect and interaction terms simultaneously by using the Group-Lasso penalty which selects or omits groups of variables when potential predictors are structured into groups known a-priori (Yuan and Lin 2011).

## 4.2 Recommendations

Future research should consider a more theoretical rigorous approach to assess prediction accuracy and variable selection of methods combining statistical learning and imputation techniques, instead of only the use of simulations.

To improve precision models of CRT and explain heterogeneity of treatment success, more modalities of data delivering sufficient phenotypic detail for the subjects, such as medical history, comorbities, progression in symptoms, genetics, neuroimaging and OMICS data, are required from a larger number of studies with a contained percentage of missingness (Bzdok and Meyer-Lindenberg 2018 and Eyre, Singh, and Reynolds 2016).

I recommend to use a MissForest-Lasso model with a tolerance penalty corresponding to an error within 3% of the minimum and not the best penalty (which gives the minimum error) to correct for the Lasso model selection inconsistency in high dimensional data (Fan and Lv 2009 and Zhao and Yu 2006). I also suggest to use this method when in presence of complete outcome, with datasets with limited percentages of missing data and mixed low and high correlations between variables, to have both good accuracy and variable selection results.

To overcome the challenge of highly correlated data with MissForest-Lasso, the performance of the Semi-standard PArtial Covariance (SPAC, Xue and Qu n.d.) method should be assessed instead of the Lasso, because it has been shown capable of diminishing correlated effects from other covariates and still incorporating signal strength. SPAC showed better variable selection performance than Lasso, adaptive Lasso and SCAD penalties.

It could be shown that the Lasso performs well for any degree of correlation if suitable tuning parameters are selected (Hebiri and Lederer 2013), but my simulation results and other studies (Fan and Lv 2009, Lu and Petkova 2014), have shown that feature variable selection may underperform for situations in which there is high correlation between active variables and noise terms. I, therefore, recommend using the Lasso method with care if variables are strongly correlated. A variable preselection for retaining only one among highly correlated variables, based on clinical expertise may be useful. Alternatively, if some strong correlation is present, the highly correlated variable should still be considered as potential true predictor in trying to understand the model.

Longitudinal analysis using statistical learning models such as penalised linear mixed models (Groll and Tutz 2012, Tutz and Groll 2011 and Schelldorfer, Meier, and Bühlmann 2014) should be used for CRT precision medicine. Accounting for the correlation between follow-up measures of the outcome and baseline and end-of-treatment at the same time in the model

might give more power to the model to capture moderators of treatment effects.

## 4.3 Concluding remarks

This PhD project simulation study showed that the combined methods MissForest-Lasso and MissForest-Conditional Random Forests have relatively good prediction accuracy and variable selection or importance performance in datasets with a high number of variables relative to sample size, with low to moderate correlations, when high percentages of missing data are present in the predictors and lower percentages of missing data are present in the outcome.

This piece of research also attempted to identify moderators of CRT with a MissForest-Lasso model applied to seven RCTs data without success, as there was no signal for moderation in the available data. Different types of data and more studies will be needed for this purpose in future research.

# References


Altman, NS (1992). ‘An introduction to kernel and Nearest-Neighbour non-parametric regression’. In: *The American Statistician* 46.3, pp. 175–185.

Andreasen, NC (1982). ‘Negative symptoms in schizophrenia: definition and reliability’. In: *Archives of General Psychiatry* 39, pp. 784–788.

— (1984). ‘Scale for Assessment of Positive symptoms’. In: *University of Iowa*.

Arbet, J et al. (2017). ‘Resampling-based tests for Lasso in genome-wide association studies’. In: *BMC Genetics* 18.70, pp. 1–15.

Arlot, S (2010). ‘A survey of cross-validation procedures for model selection’. In: *Statistics Surveys* 4, pp. 40–79.

Armitage, P, G Berry, and JNS Matthews (2002). *Statistical Methods in Medical Research*. Ed. by Blackwell Science.

Baker, FB (2001). *The basics of item response theory*. Ed. by College Park.

Barnett, JH et al. (2006). ‘Cognitive reserve in neuropsychiatry’. In: *Psychological Medicine* 36.8, pp. 1053–64.

Bartlett, MS (1937). ‘The statistical conception of mental factors’. In: *British Journal of Psychology* 2, pp. 97–104.

— (1938). ‘Methods of estimating mental factors’. In: *Nature* 141, pp. 609–610.

Bell, MD et al. (2008). ‘Neurocognitive Enhancement Therapy with Vocational Services in Schizophrenia: Work Outcomes at Two Year Follow-Up’. In: *Schizophrenia Research* 105.1-3, pp. 18–29.

Bernardini, F et al. (2017). ‘Risk Prediction Models in Psychiatry: Toward a New Frontier for the Prevention of Mental Illnesses’. In: *The Journal of Clinical Psychiatry* 78.5, pp. 575–583.

Bickel, PJ, Y Ritov, and AB Tsybakov (2009). ‘Simultaneous analysis of Lasso and Dantzig Selector’. In: *The Annals of Statistics* 37.4, pp. 1705–1732.

Bishop, CM (2007). *Pattern recognition and Machine Learning*. Ed. by Springer.

Boos, DD and LA Stefanski (2013). *Essential Statistical Inference*. Ed. by Springer.

Borra, S and A Di Ciaccio (2010). 'Repeated double cross validation'. In: *Computational Statistics & Data Analysis* 54.12, pp. 2976–2989.

Bradburn, MJ, JJ Deeks, and DG Altman (1998). 'Metan - An alternative meta-analysis command'. In: *Stata Technical Bulletin* 44, pp. 1–15.

Breiman, L (2001). 'Random Forests'. In: *Machine Learning* 45.1, pp. 5–32.

— (2003). 'Manual-setting up, using, and understanding random forests V4.0. Available at'. In: URL: https://www.stat.berkeley.edu/~breiman.

Breiman, L, A Cutler, et al. (2006). 'Breiman and Cutler's Random Forests for Classification and Regression'. In: *Machine Learning*. URL: http://CRAN.R-project.org/doc/Rnews/.

Breiman, L, JH Friedman, et al. (1984). *Classification and regression trees*. Ed. by Wadsworth & Brooks/Cole Advanced Books & Software.

Brier, GW (1950). 'Verification of forecasts expressed in terms of probability'. In: *Monthly Weather Review* 78, pp. 1–3.

Brown, TA (2006). *Confirmatory factor analysis for applied research*. Ed. by Guilford Press.

Burmann, P (1989). 'A comparative study of ordinary cross-validation,v-fold cross validation and repeating learning-testing methods'. In: *Biometrika* 76.3, pp. 503–514.

Burton, A et al. (2006). 'The design of simulation studies in medical statistics'. In: *Statistics in Medicine* 25, pp. 4279–92.

Bzdok, D and A Meyer-Lindenberg (2018). 'Machine Learning for Precision Psychiatry'. In: *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* 3, pp. 223–230.

Calaway, R et al. (2017). *doParallel: Foreach Parallel Adaptor for the 'parallel' Package*. R package version 2.14.0 or newer.

Candes, E and T Tao (2007). 'The Dantzig selector: Statistical estimation when p is much larger than n'. In: *The Annals of Statistics* 35.6, pp. 2313–2351.

Carpenter, JR and MG Kenward (2007). *Missing data in randomised controlled trials - A practical guide*. Ed. by London School of Hygiene and Tropical Medicine.

— (2013). *Multiple imputation and its application*. Ed. by Wiley.

Cella, M, AJ Bishara, et al. (2014). 'Identifying Cognitive Remediation Change Through Computational Modelling-Effects on Reinforcement Learning in Schizophrenia'. In: *Schizophrenia Bulletin* 40.6, pp. 1422–1432.

Cella, M, V Huddy, et al. (2012). 'Cognitive remediation therapy for schizophrenia'. In: *Minerva Psichiatrica* 53, pp. 185–96.

Chan, AW and DG Altman (2005). 'Epidemiology and reporting of randomised trails published in PubMed journals'. In: *The Lancet* 365, pp. 1159–1162.

Chen, Q and S Wang (2013). 'Variable selection for multiply-imputed data with application to dioxin exposure study'. In: *Statistics in Medicine* 32.21, pp. 3646–59.

*Clustering High-Dimensional Data* (2012). Naples (Italy). URL: https://sites.google.com/site/chdd12naples/.

Cortes, C and V Vapnik (1995). 'Support-Vector Networks'. In: *Machine Learning* 20, pp. 273–297.

Costello, AB and JW Osborne (2005). 'Best Practices in Exploratory Factor Analysis: Four Recommendations for Getting the Most From Your Analysis'. In: *Practical Assessment, Research & Evaluation* 10.7, pp. 1–9.

Curran, PJ and AM Hussong (June 2009). 'Integrative Data Analysis: The simultaneous analysis of multiple data sets'. In: *Psychological methods* 14.2, pp. 81–100.

Cutler, A et al. (2009). *High-dimensional data analysis in cancer research*. Ed. by Springer.

Debray, TPA et al. (2015). 'A new framework to enhance the interpretation of external validation studies of clinical prediction models'. In: *Journal of Clinical Epidemiology* 68.3, pp. 279–289.

Dempster, AP, NM Laird, and DB Rubin (1977). 'Maximum likelihood from incomplete data via the EM algorithm'. In: *Journal of the Royal Statistical Society* 39.1, pp. 1–38.

Dunn, G et al. (2015). 'Evaluation and validation of social and psychological markers in randomised trials of complex interventions in mental health: a methodological research programme'. In: *Health Technology Assessment* 19.93.

Dwyer, BD, P Falkai, and N Koutsouleris (2018). 'Machine Learning Approaches for Clinical Psychology and Psychiatry'. In: *Annual Review of Clinical Psychology and Psychiatry* 14, pp. 12.1–12.28.

Eack, SM et al. (2009). 'Cognitive Enhancement Therapy for Early-Course Schizophrenia: Effects of a Two-Year Randomized Controlled Trial'. In: *Psychiatric Services* 60.11, pp. 1468–76.

Efron, B (1979). 'Bootstrap methods: Another look at the Jackknife'. In: *The Annals of Statistics* 7.1, pp. 1–16.

Efron, B and R Tibshirani (1997). 'Improvements on cross-validation: the .632+ bootstrap method'. In: *Journal of the American Statistical Association* 92.438, pp. 548–560.

Efron, B and RJ Tibshirani (1994). *An introduction to the Bootstrap*. Ed. by Chapman & Hall/CRC.

Estabrook, R and M Neale (2013). 'A comparison of factor score estimation methods in the presence of missing data: reliability and an application to nicotine dependence'. In: *Preactical Assessment, Research & Evaluation* 48.1, pp. 1–27.

Everitt, BS and S Wessely (2009). *Clinical Trials in Psychiatry*. Ed. by Wiley.

Eyre, HA, AB Singh, and C Reynolds (2016). 'Tech giants enter mental health'. In: *World Psychiatry* 15, pp. 21–22.

Fabrigar, LR et al. (1999). 'Evaluating the Use of Exploratory Factor Analysis in Psychological Research'. In: *Psychological Methods* 4.3, pp. 272–299.

Fan, J and R Li (2001). 'Variable selection via nonconcave penalized likelihood and its oracle properties'. In: *Journal of the American Statistical Association* 96, pp. 1348–1360.

Fan, J and J Lv (2009). 'A selective overview of variable selection in high dimensional feature space'. In: *Statistica Sinica* 20, pp. 101–148.

Fioravanti, M et al. (2005). 'A meta-analysis of cognitive deficits in adults with a diagnosis of schizophrenia'. In: *Neuropsychology Review* 15.2, pp. 73–95.

Fiszdon, JM et al. (2016). 'Cognitive remediation for individuals with psychosis: efficacy and mechanisms of treatment effects'. In: *Psychological Medicine* 46, pp. 3275–3289.

Friedman, J, T Hastie, and R Tibshirani (2010). 'glmnet: Regularisation Paths for Generalized Linear Models via Coordinate Descent'. In: 33.1, pp. 1–22. URL: http://www.jstatsoft.org/v33/i01/.

Friedman, JH (1991). 'Multivariate adaptive regression splines'. In: *The Annals of Statistics* 1, pp. 1–141.

Friedman, N, D Geiger, and M Goldszmidt (1997). 'Bayesian Network Classifiers'. In: *Machine Learning* 29.2-3, pp. 131–163.

Garety, PA et al. (2008). 'Cognitive-behavioural therapy and family intervention for relapse prevention and symptom reduction in psychosis'. In: *The British Journal of Psychiatry* 192, pp. 412–423.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. MIT Press. URL: http://www.deeplearningbook.org.

Green, MF and PD Harvey (2014). 'Cognition in schizophrenia: past, present and future'. In: *Schizophrenia Research: Cognition* 1.1, e1–e9.

Groll, A and G Tutz (2012). 'Variable selection for generalized linear mixed models by L1-penalized estimation'. In: *Statistics and Computing* 24.2, pp. 137–154.

Hahn, T, AA Nierenberg, and S Whitfield-Gabrieli (2017). 'Predictive analytics in mental health: applications, guidelines, challenges and perspectives'. In: *Molecular Psychiatry* 22, pp. 37–43.

Hand, DJ (2006). 'Classifier technology and the illusion of progress'. In: *Statistical Science* 21.1, pp. 1–14.

Hardt, J, M Herke, and R Leonhart (2012). 'Auxiliary variables in multiple imputation in regression with missing X: a warning against including too many in small sample research'. In: *BMC Medical Research Methodology* 12.184.

Harrell, FE (2001). *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Ed. by Springer Science & Business Media.

Harrell, FE, KL Lee, and DB Mark (1996). 'Multivariate prognostic models: issues in developing models, evaluating assumptions and accuracy, and measuring and reproducing errors'. In: *Stat Med* 15, pp. 361–387.

Harrison, RL (2010). 'Introduction to Monte Carlo Simulation'. In: *AIP Conf Proc.* 1204, pp. 17–21.

Hastie, T (2015). *Statistical learning with Sparsity: The Lasso and the generalizations*. Ed. by Chapman & Hall. Chap. 6.

Hastie, T, R Tibshirani, and J Friedman (2008). *The elements of Statistical Learning: data Mining, inference, and prediction*. Ed. by Springer.

Hebiri, M and JC Lederer (2013). 'How Correlations Influence Lasso Prediction'. In: *IEEE Transactions on Information Theory* 59.3. URL: https://arxiv.org/pdf/1204.1605.pdf.

*Help me understanding genetics: Precision medicine* (2018). URL: https://ghr.nlm.nih.gov/primer/precisionmedicine.

Heymans, MW et al. (2007). 'Variable selection under multiple imputation using the bootstrap in a prognostic study'. In: *BMC Medical Research Methodology* 7, pp. 33–42.

Higgins, JP et al. (Aug. 2001). 'Meta-analysis of continuous outcome data from individual patients'. In: *Stat Med.* 20.15, pp. 2219–41.

Hingorani, AD et al. (2013). 'Prognosis research strategy (PROGRESS) 4: Prognostic factor research'. In: *PLoS Med*. URL: doi:10.1136/bmj.e5793.

Hippel, P von and J Lynch (2013). 'Efficiency Gains from Using Auxiliary Variables in Imputation'. In: URL: arXiv:1311.5249.

Hothorn, T, P B uhlmann, et al. (2006). 'Survival Ensembles'. In: *Biostatistics* 7.3, pp. 355–373.

Hothorn, T, K Hornik, and A Zeileis (2006). 'Unbiased recursive partitioning: a conditional inference framework'. In: *Journal of Computational and Graphical Statistics* 15, pp. 651–674.

Ibrahim, JG, SR Lipsitz, and N Horton (2001). 'Using auxiliary data for parameter estimation with non-ignorably missing outcomes'. In: *Applied Statistics* 50, pp. 361–73.

Iniesta, R, D Stahl, and P McGuffin (2016). 'Machine learning, statistical learning and the future of biological research in psychiatry'. In: *Psychological Medicine* 46, pp. 2455–2465.

Ioannidis, JPA (2005). 'Why Most Published Research Findings Are False'. In: *PLoS Med.* 2.8, e124.

Ishwaran, H and UB Kogalur (2016). 'randomForestSRC: Random Forests for Survival, Regression and Classification (RF-SRC). R package version 2.0.5'. In: URL: http://cran.r-project.org.

Ishwaran, H, UB Kogalur, et al. (2008). 'Random survival forests'. In: *Annals of Applied Statistics* 2, pp. 841–860.

Jaaskelainen, E et al. (2013). 'A systematic review and meta-analysis of recovery in schizophrenia'. In: *Schizophr Bull.* 39.6, pp. 1296–306.

Jacobucci, R, KJ Grimm, and JJ McArdle (2016). 'Regularized Structural Equation Modeling'. In: *Structural Equation Modelling* 23.4, pp. 555–566.

James, G et al. (2013). *An introduction to Statistical Learning with applications in R*. Ed. by Springer.

Jönsson, P and C Wohlin (2004). 'An evaluation of k-nearest neighbour imputation using Likert data'. In: *Proc. 10th International Symposium pn Software Metrics*, pp. 108–118.

Kay, SR, A Fiszbein, and LA Opler (1987). 'The Positive and Negative Syndrome Scale (PANSS) for schizophrenia'. In: *Schizophrenia Bulletin* 13, pp. 261–276.

Keefe, RS et al. (2012). 'Feasibility and pilot efficacy results from the multisite cognitive remediation in the schizophrenia trials network (CRSTN) randomized controlled trial'. In: *Journal of Clinical Psychiatry* 73.7, pp. 1016–22.

Keshavan, MS et al. (2008). 'Schizophrenia, "just the facts": what we know in 2008. Part 3: Neurobiology'. In: *Schizophr Res.* 106.2-3, pp. 89–107.

Khondoker, M et al. (2013). 'A comparison of machine learning methods for classification using simulation with multiple real data examples from mental health studies'. In: *Statistical Methods in Medical Research* 25.5, pp. 1804–1823.

Koenen, KC et al. (2009). 'Childhood IQ and adult mental disorders: a test of the cognitive reserve hypothesis'. In: *The American Journal of Psychiatry* 166.1, pp. 50–57.

König, IR et al. (2007). 'Practical experiences on the necessity of external validation'. In: *Statistics in Medicine* 26, pp. 5499–5511.

Koutsouleris, N et al. (2016). 'Multisite prediction of 4-week and 52-week treatment outcomes in patients with first-episode psychosis: a machine learning approach'. In: *Lancet Psychiatry* 3.10, pp. 935–946.

Kraemer, HC, E Frank, and DJ Kupfer (2006). 'Moderators of Treatment Outcomes: Clinical, Research, and Policy Importance'. In: *JAMA* 296, pp. 1286–1289.

Kraemer, HC, GT Wilson, et al. (2002). 'Mediators and Moderators in Randomized Clinical Trials'. In: *Arch Gen Psychiatry* 59, pp. 877–83.

Kriesel, David (2007). *A Brief Introduction to Neural Networks*. URL: available%20at%20http://www.dkriesel.com.

Krstajic, D et al. (2014). 'Cross-validation pitfalls when selecting and assessing regression and classification models'. In: *Journal of Cheminformatics* 6.1, p. 10.

Kuhn, M (2016). *caret: Classification and Regression Training*. R package version 2.10 or newer.

Kuhn, M and K Johnson (2013). *Applied Predictive Modeling*. Ed. by Springer.

Kupper, LL and MD Hogan (1978). 'Interaction in epidemiologic studies'. In: *American Journal of Epidemiology* 108, pp. 447–453.

Kurtz, MM et al. (2009). 'Predictors of Change in Life Skills in Schizophrenia after Cognitive Remediation'. In: *Schizophrenia Research* 107.2-3, pp. 267–274.

Lattin, J, D Carroll, and PE Green (2003). *Analyzing multivariate data*. Ed. by Thomson & Brooks/Cole.

Leon, J de (2012). 'Evidence-Based Medicine versus Personalized Medicine: Are They Enemies?' In: *Journal of Clinical Psychopharmacology* 32.2, pp. 153–164. URL: doi:%2010.1097/JCP.0b013e3182491383.

Liao, SG et al. (2014). 'Missing value imputation in high-dimensional phenomic data: imputable or not, and how?' In: *BMC Bioinformatics* 15, p. 346.

Liaw, A and M Wiener (2002a). 'Classification and Regression by randomForest'. In: *R News* 2.3, pp. 18–22. URL: https://CRAN.R-project.org/doc/Rnews/.

— (2002b). 'Classification and regression by randomForest'. In: *Rnews* 2.3, pp. 18–22.

Lindenmayer, JP, R Bernstein-Hyman, and S Grochowski (1994). 'A new five factor model of schizophrenia'. In: *Psychiatric Quaterly* 65, pp. 299–322.

Lindenmayer, JP, VA Ozog, et al. (2017). 'Predictors of response to cognitive remediation in service recipients with severe mental illness'. In: *Psychiatric Rehabilitation Journal* 40.1, pp. 61–69.

Liu, BY et al. (2016). ‘Variable selection and prediction with incomplete high-dimensional data’. In: *Annals of Applied Statistics* 10.1, pp. 418–450.

Lockhart, R et al. (2014). ‘A significance test for the lasso’. In: *The Annals of Statistics* 42.2, pp. 413–468.

Lu, F and E Petkova (2014). ‘A comparative study of variable selection methods in the context of developing psychiatric screening instruments’. In: *Statistics in Medicine* 33.3, pp. 401–21.

McGurk, SR et al. (2007). ‘A meta-analysis of Cognitive remediation in Schizophrenia’. In: *The American Journal of Psychiatry* 164.12, pp. 1791–1802.

Meredith, W and JA Teresi (2006). ‘An Essay on Measurement and Factorial Invariance’. In: *Medical Care* 44.11, S69–S77.

Moons, KG et al. (2006). ‘Using the outcome of missing predictor values was preferred’. In: *Journal of Clinical Epidemiology* 59.10, pp. 1092–101.

Moritz, S et al. (2015). ‘Comparison of different Methods for Univariate Time Series Imputation in R’. In: URL: ArXiv%20e-prints.

Morrison, A (2001). ‘The interpretation of intrusions in psychosis: an integrative approach to hallucinations and delusions’. In: *Behav. Cogn. Psychoter.* 29, pp. 257–276.

Musoro, JZ et al. (2014). ‘Validation of prediction models based on lasso regression with multiply imputed data’. In: *BMC Medical Research Methodology* 14, p. 116.

Muthén, BO and LK Muthén (n.d.). *Factor scores*. http://www.statmodel.com/discussion/messages/9/8465.html?1439173196.

Muthén, BO and J Yang Hsu (1993). ‘Selection and predictive validity with latent variable structures’. In: *British Journal of Mathematical and Statistical Psychology* 46, pp. 255–271.

Muthén, LK and BO Muthén (2012). *Mplus User’s Guide. Seventh Edition*. Ed. by CA: Muthén & Muthén Los Angeles.

Niloofar, P and M Ganjali (2014). ‘A new multivariate imputation method based on Bayesian networks’. In: *Journal of Applied Statistics* 41.3, pp. 501–518.

Nuzzo, R (2014). ‘STATISTICAL ERRORS: P values, the “gold standard” of statistical validity, are not as reliable as many scientists assume’. In: *Nature* 506, pp. 150–153.

Oba, S et al. (2003). ‘A Bayesian missing value estimation method for gene expression profile data’. In: *Bioinformatics* 19.16, pp. 2088–2096.

Overall, JE and DR Gorham (1962). ‘The Brief Psychiatric Rating Scale’. In: *Psychological Reports* 10, pp. 799–812.

Park, T and G Casella (2008). ‘The Bayesian Lasso’. In: *Journal of the American Statistical Association* 103, pp. 681–686.

Quirk, A et al. (2013). 'Development of the carer well-being and support (CWS) questionnaire'. In: *Mental Health Review Journal* 17.3, pp. 128–138.

Rajji, TK, MA Miranda, and BH Mulsant (2014). 'Cognition, function, and disability in patients with schizophrenia: a review of longitudinal studies'. In: *Can J Psychiatry* 59.1, pp. 13–17.

Ramsay, IS et al. (2018). 'Model selection and predictions of outcomes in recent onset schizophrenia patients who undergo cognitive training'. In: *Schizophrenia Research: Cognition* 11, pp. 1–5.

Redekop, WK and D Mladsi (2013). 'The Faces of Personalized Medicine: A Framework for Understanding Its Meaning and Scope'. In: *Elsevier* 16, pp. 54–59.

Reeder, C et al. (2017). 'A new generation computerised metacognitive cognitive remediation programme for schizophrenia (CIRCuiTS): a randomised controlled trial'. In: *Psychological medicine*.

Riley, RD et al. (2013). 'Prognosis research strategy (PROGRESS) 2: Prognostic factor research'. In: *PLoS Med*. URL: doi:10.1371.journal/pmed.1001380.

Rosseel, Y (2012). 'lavaan: An R Package for Structural Equation Modeling'. In: *Journal of Statistical Software* 48.2, pp. 1–36. URL: http://www.jstatsoft.org/v48/i02/.

Rubin, DB (1976). 'Inference and missing data'. In: *Biometrika* 3, pp. 581–92.

— (1981). *Multiple Imputation for Nonresponse in Surveys*. Ed. by John Wiley & Sons.

Rubin, DB, HS Stern, and V Hehovar (1995). 'Handling 'don't know' survey responses: the case of the Slovenian plebiscite'. In: *Journal of the American Statistical Association* 90, pp. 822–8.

Schelldorfer, J, L Meier, and P Bühlmann (2014). 'GLMMLasso: An algorithm for high-dimensional generalized linear mixed models using $l_1$-penalization'. In: *Journal of Computational and Graphical Statistics* 23.2, pp. 460–477.

Seixas, FL et al. (2014). 'A Bayesian network decision model for supporting the diagnosis of dementia, Alzheimer's disease and mild cognitive impairment'. In: *Computers in Biology and Medicine* 51, pp. 140–158.

Shah, AD et al. (2014). 'Comparison of Random Forest and Parametric Imputation Models for Imputing Missing Data using MICE: a CALIBER Study'. In: *Am J Epidemiol.* 179.6, pp. 764–774.

Shao, J and RR Sitter (1996). 'Bootstrap for imputed survey data'. In: *Journal of the American Statistical Association* 91.435, pp. 1278–1288.

Shmueli, G (2010). 'To Explain or to Predict?' In: *Statistical Science* 25.3, pp. 289–310. URL: doi:%2010.1214/10-STS330.

Sibbald, B and M Roland (1998). 'Understanding controlled trials. Why are randomised controlled trials important?' In: *BMJ* 316.7126, p. 201.

Silverstein, SM et al. (2009). 'Attention shaping: a reward-based learning method to enhance skills training outcomes in schizophrenia'. In: *Schizophrenia Bulletin* 35.1, pp. 222–232.

Skrondal, A (2001). 'Regression among factor scores'. In: *Psychometrika* 66, pp. 563–575.

SL, Hershberger (2005). 'Factor scores'. In: *Encyclopedia of Statistics in Behavioral Science*. Ed. by B. S. Everitt and D. C. Howell. New York: John Wiley, pp. 636–644.

Smith, GCS et al. (2014). 'Correcting for Optimistic Prediction in Small Data Sets'. In: *Am J Epidemiol* 180.3, pp. 318–324.

Smola, AJ and B Schölkopf (1998). 'A tutorial on Support Vector Regression'. In: *NeuroCOLT2 Technical Report Series*.

Stahl, D and A Pickles (2018). 'Fact or fiction: reducing the proportion and impact of false positives'. In: *Psychological Medicine* 48, pp. 1084–1091.

Stahl, D, A Pickles, et al. (2012). 'Novel Machine Learning Methods for ERP Analysis: A Validation From Research on Infants at Risk for Autism'. In: *Developmental Neuropsychology* 37.3, pp. 274–298.

Stata (2015). *Stata Statistical Software: Release 14*. Version 14.

Stekhoven, DJ (2013). *missForest: Nonparametric Missing Value Imputation using Random Forest*. R package version 1.4.

Stekhoven, DJ and P Buhlmann (2012). 'MissForest - non-parametric missing value imputation for mixed-type data'. In: *Bioinformatics* 28.1, pp. 112–118.

Sterne, JAC et al. (2009). 'Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls'. In: *BMJ* 338. URL: http://dx.doi.org/10.1136%2Fbmj.b2393.

Steyerberg, EW (2009). *Clinical Prediction Models: a practical approach to development, validation, and updating*. Ed. by Springer.

Steyerberg, EW and FE Harrell (2016). 'Prediction models need appropriate internal, internal-external, and external validation'. In: *Journal of Clinical Epidemiology* 69, pp. 245–247.

Steyerberg, EW and Y Vergouwe (2014). 'Towards better clinical prediction models: seven steps for development and an ABCD for validation'. In: *European Heart Journal* 35, pp. 1925–1931.

Strobl, C, AL Boulesteix, and A Zeileis (2007). 'Bias in random forest variable importance measures: illustrations, sources and a solution'. In: *BMC Bioinformatics* 8.25.

Strobl, C, A Boulesteix, et al. (2008). 'Conditional variable importance for Random Forests'. In: *BMC Bioinformatics* 9, p. 307.

Tang, F and H Ishwaran (2017). 'Random Forest Missing Data Algorithms'. In: *Bioinformatics*. URL: arXiv:1701.05305v2%20[stat.ML].

Tawfik, AY and SD Goodwin (2004). *Advances in Artificial Intelligence*. Ed. by Springer-Verlag Berlin Heidelberg.

Team, R Core (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Teixeira-Pinto, A et al. (July 2009). 'Statistical Approaches to Modeling Multiple Outcomes In Psychiatric Studies'. In: *Psychiatr Ann.* 39.7, pp. 729–735.

Thomson, GH (1934). 'The meaning of 'i' in the estimate of 'g''. In: *British Journal of Psychology* 25, pp. 92–99.

Thurston, LL (1935). *The vectors of mind*. Ed. by University of Chicago Press.

Tibshirani, R (1996). 'Regression shrinkage and selection via the lasso'. In: *Journal of the Royal Statistical Society, Series B* 58.1, pp. 267–288.

Troyanskaya, O. et al. (2001). 'Missing value estimation methods for DNA microarrays'. In: *Bioinformatics* 17, pp. 520–525.

Tutz, G and A Groll (2011). 'Variable selection for generalized linear mixed models by L1-penalized estimation'. In: *Statistics and Computing* 24.2, pp. 137–154.

Van Buuren, S (2007). 'Multiple imputation of discrete and continuous data by fully conditional specification'. In: *Statistical methods in Medical Research* 16.3, pp. 219–242.

Van Buuren, S and K Groothuis-Oudshoorn (2011). 'mice: Multivariate Imputation by Chained Equations in R'. In: *Journal of Statistical Software* 45.3, pp. 1–67. URL: http://www.jstatsoft.org/v45/i03/.

Van Buuren, S and CGM Oudshoorn (2000). 'Multivariate imputation by chained equations: MICE V1.0 User's manual'. In: *TNO Report PG/VGZ/00.038*. URL: http://www.multiple-imputation.com/.

Van Os, J and S Kapur (2009). 'Schizophrenia'. In: *Lancet* 374, pp. 635–45.

VanderWeele, TJ (2009). 'On the distinction between interaction and effect modification'. In: *Epidemiology* 20, pp. 863–871.

— (2015). *Explanation in Causal Inference: Methods for Mediation and Interaction*. Ed. by Oxford University Press.

Waljee, A et al. (2013). 'Comparison of imputation methods for missing laboratory data in medicine'. In: *BMJ Open* 3.8. URL: e002847.

Wan, Y et al. (2015). 'Variable selection models based on multiple imputation with an application for predicting median effective dose and maximum effect'. In: *Journal of Statistical Computing and Simulation* 85.9, pp. 1902–1916.

Wium-Andersen, IK et al. (2016). 'Personalized medicine in psychiatry'. In: *Nordic Journal of Psychiatry* 71.1, pp. 12–19.

Wood, AM, IR White, and P Royston (2008). 'How should variable selection be performed with multiply imputed data?' In: *Statistics in Medicine* 27, pp. 3227–3246.

Wood, AM, IR White, and SG Thompson (2004). 'Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals'. In: *Clinical Trials* 1, pp. 368–376.

Wykes, T, M Brammer, et al. (2002). 'Effects on the brain of a psychological treatment: cognitive remediation therapy. Functional magnetic resonance imaging in schizophrenia'. In: *The British Journal of Psychiatry* 181.2, pp. 144–152.

Wykes, T, V Huddy, et al. (2011). 'A meta-analysis of cognitive remediation for schizophrenia: methodologies and effect sizes'. In: *The American Journal of Psychiatry* 12, pp. 416–21.

Wykes, T, E Newton, et al. (2007). 'Cognitive remediation therapy (CRT) for young early onset patients with schizophrenia: An exploratory randomized controlled trial'. In: *Schizophrenia Research* 94.1-3, pp. 221–30.

Wykes, T, C Reeder, J Corner, et al. (1999). 'The Effects of Neurocognitive Remediation on Executive Processing in Patients With Schizophrenia'. In: *Schizophrenia Bulletin* 2, pp. 291–307.

Wykes, T, C Reeder, S Landau, B Everitt, et al. (2007). 'Cognitive remediation therapy in schizophrenia: Randomised controlled trial'. In: *British Journal of Psychiatry* 190.5, pp. 421–427.

Wykes, T, C Reeder, S Landau, P Matthiasson, et al. (2009). 'Does age matter? Effects of cognitive rehabilitation across the age span'. In: *Schizophrenia Research* 113, pp. 252–258.

Xue, F and A Qu (n.d.). 'Variable Selection for Highly Correlated Predictors'. In: (). URL: arXiv:1709.04840%20[stat.ME].

Yeo, IK and RA Johnson (2000). 'A new family of power transformations to improve normality or symmetry'. In: *Biometrika* 87.4, pp. 954–959.

Yuan, M and Y Lin (2011). 'Model Selection and Estimation in Regression with Grouped Variables'. In: *Journal of the Royal Statistical Society, Series B* 68, pp. 49–67.

Yunus, M (2017). ‘Characteristics of Group Lasso in handling high correlated data’. In: *Applied Mathematical Sciences* 20, pp. 953–961.

Zhao, P and B Yu (2006). ‘On model selection consistency of Lasso’. In: *Journal of Machine Learning Research* 7, pp. 2541–2563.

Zou, H (2006). ‘The Adaptive Lasso and its oracle properties’. In: *Journal of the American Statistical Association* 101.476, pp. 1418–1429.

Zou, H and T Hastie (2005). ‘Regularization and Variable Selection via the Elastic Net’. In: *Journal of the Royal Statistical Society, Series B* 67.2, pp. 301–320.

# Appendices

# Appendix A

# Other simulation result tables and figures

## A.1 MICE-Lasso and MICE-Elasticnet simulations results tables and figures

Table A.1: **Accuracy** simulation study results for **MICE-Lasso** analysis with **bootstrap validation on the remaining data** for scenarios **S1-S2**, based on 300 data sets of **20 variables** each (**n=250**). Means of all estimates along with their corresponding 2.5th and 97.5th percentile values within parenthesis. The theoretical MSE is $1.74^2 = 3.028$

| **Estimates** | **Complete cases** | | | **MCAR** | | | **MAR** | | |
|---|---|---|---|---|---|---|---|---|---|
| | Best model | 10% tolerance | 40% tolerance | Best model | 10% tolerance | 40% tolerance | Best model | 10% tolerance | 40% tolerance |
| $\text{MSE}_{internal}$ | 3.479 (3.236,3.710) | 3.593 (3.345,3.826) | 4.463 (4.158,4.763) | 4.721 (4.390,5.050) | 4.348 (4.048,4.614) | 4.530 (4.185,4.833) | 4.383 (4.071,4.642) | 4.186 (3.893,4.424) | 4.593 (4.255,4.912) |
| $\beta_{LP^*}$ | 1.002 (0.993,1.010) | 1.102 (1.090,1.115) | 1.290 (1.273,1.309) | 0.871 (0.849,0.891) | 0.970 (0.948,0.993) | 1.175 (1.149,1.205) | 0.918 (0.902,0.933) | 1.016 (0.996,1.035) | 1.218 (1.197,1.242) |
| $\alpha_{LP^*}$ | -0.006 (-0.030, 0.016) | -0.195 (-0.229,-0.160) | -0.548 (-0.602,-0.489) | 0.215 (0.167,0.259) | 0.031 (-0.022,0.081) | -0.349 (-0.408,-0.287) | 0.130 (0.088,0.173) | -0.054 (-0.105,-0.004) | -0.433 (-0.501,-0.368) |

Table A.2: **Accuracy** simulation study results for **MICE-Lasso** analysis with **bootstrap validation on the remaining data** for scenarios **S1-S2** based on 300 data sets of **20 variables** each (**n=1000**). Means of all estimates along with their corresponding 2.5th and 97.5th percentile values within parenthesis. The theoretical MSE is $1.74^2 = 3.028$

| **Estimates** | **Complete cases** | | | **MCAR** | | | **MAR** | | |
|---|---|---|---|---|---|---|---|---|---|
| | Best model | 10% tolerance | 40% tolerance | Best model | 10% tolerance | 40% tolerance | Best model | 10% tolerance | 40% tolerance |
| $\text{MSE}_{internal}$ | 3.137 (3.032,3.245) | 3.182 (3.081,3.291) | 4.125 (3.991,4.269) | 3.361 (3.236,3.482) | 3.342 (3.215,3.469) | 4.111 (3.944,4.272) | 3.366 (3.243,3.486) | 3.360 (3.232,3.481) | 4.168 (4.010,4.323) |
| $\beta_{LP^*}$ | 1.014 (1.010,1.017) | 1.074 (1.069,1.080) | 1.309 (1.301,1.319) | 0.986 (0.982,0.990) | 1.041 (1.035,1.047) | 1.274 (1.263,1.283) | 0.993 (0.989,0.997) | 1.048 (1.042,1.054) | 1.285 (1.274,1.294) |
| $\alpha_{LP^*}$ | -0.027 (-0.038, 0.017) | -0.141 (-0.155,-0.127) | -0.584 (-0.613,-0.556) | 0.020 (0.008,0.032) | -0.084 (-0.097,-0.072) | -0.525 (-0.551,-0.492) | 0.014 (-0.000,0.028) | -0.092 (-0.108,-0.076) | -0.542 (-0.570,-0.512) |

Table A.3: **Accuracy** simulation study results for **MICE-Lasso** analysis with Harrell (1996) bootstrap validation: scenarios **S1** (without missing data, no assumption of moderation) and **S2** (with missing data, complete outcome, no assumption of moderation) based on 300 data sets of **20 variables** each (**n=1000**). Means of all estimates along with their corresponding 2.5th and 97.5th percentile values within parenthesis. The theoretical MSE is $1.74^2 = 3.028$

| **Estimates** | **Complete cases** | | | |
|---|---|---|---|---|
| | Best model | 1 SE tolerance | 3% tolerance | 15% tolerance |
| $\text{MSE}_{apparent}$ | 2.985 (2.718,3.274) | 3.024 (2.749,3.317) | 3.189 (2.887,3.509) | 3.922 (3.542,4.295) |
| $\beta_{LP}$ | 1.032 (1.022,1.041) | 1.060 (1.050,1.072) | 1.137 (1.118,1.156) | 1.309 (1.263,1.354) |
| Tuning $\lambda$ | 0.034 (0.022,0.043) | 0.064 (0.054,0.076) | 0.144 (0.134,0.149) | 0.324 (0.293,0.367) |
| $\text{MSE}_{ext}$ | 3.144 (3.505,3.757) | 3.176 (3.543,3.795) | 3.369 (3.256, 3.491) | 4.244 (3.958,4.499) |
| $\text{Optimism}_{ext}$ | -0.159 (-0.433,0.149) | -0.152 (-0.424,0.148) | -0.180 (-0.480,0.123) | -0.323 (-0.687,0.049) |
| $\text{Optimism}_{int}$ | -0.107 (-0.136,-0.077) | -0.074 (-0.120,-0.064) | -0.074 (-0.101,-0.045) | -0.077 (-0.112,-0.044) |
| $\text{MSE}_{corrected}$ | 3.092 (2.819,3.397) | 3.116 (2.836,3.418) | 3.264 (2.996,3.594) | 3.998 (3.616,4.357) |
| $\beta_{LP^*}$ | 1.020 (1.014,1.025) | 1.049 (1.040,1.056) | 1.126 (1.111,1.141) | 1.297 (1.255,1.335) |
| | | **MCAR** | | |
| $\text{MSE}_{apparent}$ | 2.999 (2.650,3.398) | 3.032 (2.682,3.437) | 3.201 (2.823,3.630) | 3.932 (3.486,4.443) |
| $\beta_{LP}$ | 1.023 (1.016,1.031) | 1.051 (1.042,1.062) | 1.128 (1.113,1.148) | 1.301 (1.261,1.345) |
| Tuning $\lambda$ | 0.026 (0.017,0.034) | 0.055 (0.042,0.065) | 0.135 (0.120,0.150) | 0.317 (0.287,0.351) |
| $\text{MSE}_{ext}$ | 3.179 (3.107,3.284) | 3.194 (3.111,3.303) | 3.369 (3.230,3.552) | 4.274 (3.943,4.686) |
| $\text{Optimism}_{ext}$ | -0.181 (-0.527,0.192) | -0.162 (-0.506,0.192) | -0.168 (-0.509,0.182) | -0.315 (-0.673,0.081) |
| $\text{Optimism}_{int}$ | -0.179 (-0.256,-0.108) | -0.156 (-0.231,-0.089) | -0.115 (-0.184,-0.054) | -0.097 (-0.147,-0.043) |
| $\text{MSE}_{corrected}$ | 3.178 (2.819,3.605) | 3.188 (2.830,3.617) | 3.315 (2.935,3.756) | 4.028 (3.588,4.553) |
| $\beta_{LP^*}$ | 1.005 (0.994,1.016) | 1.033 (1.020,1.047) | 1.110 (1.092,1.129) | 1.281 (1.240,1.323) |
| | | **MAR** | | |
| $\text{MSE}_{apparent}$ | 2.993 (2.654,3.324) | 3.028 (2.686,3.363) | 3.196 (2.842,3.551) | 3.923 (3.477,4.351) |
| $\beta_{LP}$ | 1.025 (1.017,1.032) | 1.053 (1.042,1.063) | 1.130 (1.114,1.147) | 1.303 (1.263,1.343) |
| Tuning $\lambda$ | 0.027 (0.018,0.035) | 0.057 (0.045,0.066) | 0.137 (0.123,0.151) | 0.318 (0.290,0.343) |
| $\text{MSE}_{ext}$ | 3.172 (3.101,3.265) | 3.191 (3.111,3.292) | 3.369 (3.241,3.538) | 4.241 (3.918,4.605) |
| $\text{Optimism}_{ext}$ | -0.179 (-0.508,0.150) | -0.163 (-0.501,0.149) | -0.174 (-0.510,0.132) | -0.318 (-0.706,0.066) |
| $\text{Optimism}_{int}$ | -0.169 (-0.205,-0.138) | -0.143 (-0.178,-0.110) | -0.114 (-0.145,-0.084) | -0.115 (-0.148,-0.077) |
| $\text{MSE}_{corrected}$ | 3.162 (2.819,3.504) | 3.171 (2.825,3.519) | 3.310 (2.942,3.667) | 4.037 (3.591,4.474) |
| $\beta_{LP^*}$ | 1.019 (1.012,1.025) | 1.038 (1.029,1.046) | 1.097 (1.085,1.108) | 1.228 (1.200,1.259) |

Table A.4: **Accuracy** simulation study results for **MICE-Elasticnet** analysis with Harrell bootstrap validation: scenarios **S1** (without missing data, no assumption of moderation) and **S2** (with missing data, complete outcome, no assumption of moderation) based on 300 data sets of **20 variables** each (**n=1000**). Means of all estimates along with their corresponding 2.5th and 97.5th percentile values within parenthesis. The theoretical MSE is $1.74^2 = 3.028$

| **Estimates** | **Complete cases** | | | |
|---|---|---|---|---|
| | Best model | 1 SE tolerance | 3% tolerance | 15% tolerance |
| MSE$_{apparent}$ | 2.985 (2.716,3.270) | 3.216 (2.900,3.545) | 4.985 (4.031,5.513) | 7.741 (6.571,8.406) |
| $\beta_{LP}$ | 1.032 (1.024,1.043) | 1.147 (1.111,1.199) | 1.451 (1.356,1.542) | 2.785 (1.981,3.406) |
| Tuning $\alpha$ | 0.880 (0.700,0.900) | 0.900 (0.900,0.900) | 0.900 (0.900,0.900) | 0.900 (0.900,0.900) |
| Tuning $\lambda$ | 0.038 (0.032,0.051) | 0.170 (0.135,0.220) | 0.558 (0.455,0.580) | 1.487 (1.199,1.528) |
| MSE$_{ext}$ | 3.144 (3.096,3.206) | 3.402 (3.233,3.645) | 5.491 (4.479,6.069) | 8.580 (7.621,9.216) |
| Optimism$_{ext}$ | -0.159 (-0.434,0.149) | -0.186 (-0.479,0.113) | -0.505 (-0.995,-0.044) | -0.839 (-1.485,-0.180) |
| Optimism$_{int}$ | -0.108 (-0.136,-0.077) | -0.077 (-0.105,-0.047) | -0.076 (-0.118,-0.038) | -0.035 (-0.098,0.027) |
| MSE$_{corrected}$ | 3.093 (2.819,3.393) | 3.293 (2.973,3.636) | 5.061 (4.100,5.591) | 7.776 (6.578,8.416) |
| $\beta_{LP^*}$ | 1.021 (1.015,1.026) | 1.134 (1.105,1.158) | 1.396 (1.301,1.470) | 2.399 (1.916,2.945) |
| | **MCAR** | | | |
| MSE$_{apparent}$ | 2.999 (2.650,3.398) | 3.199 (2.762,3.630) | 4.408 (3.389,5.318) | 6.949 (4.889,8.163) |
| $\beta_{LP}$ | 1.024 (1.018,1.032) | 1.134 (1.086,1.177) | 1.397 (1.255,1.496) | 2.312 (1.703,3.045) |
| Tuning $\alpha$ | 0.727 (0.380,0.895) | 0.899 (0.880,0.900) | 0.900 (0.900,0.900) | 0.900 (0.900,0.900) |
| Tuning $\lambda$ | 0.035 (0.032,0.045) | 0.176 (0.147,0.214) | 0.525 (0.455,0.580) | 1.417 (1.199,1.570) |
| MSE$_{ext}$ | 3.180 (3.108,3.285) | 3.378 (3.190,3.636) | 4.814 (3.683,5.805) | 7.690 (5.577,8.930) |
| Optimism$_{ext}$ | -0.182 (-0.528,0.191) | -0.178 (-0.513,0.168) | -0.406 (-0.877,0.049) | -0.742 (-1.350,-0.129) |
| Optimism$_{int}$ | -0.181 (-0.258,-0.110) | -0.136 (-0.215,-0.075) | -0.104 (-0.157,-0.052) | -0.061 (-0.121,-0.007) |
| MSE$_{corrected}$ | 3.020 (2.484,3.611) | 3.496 (2.712,4.345) | 4.555 (2.936,6.167) | 6.919 (3.743,9,497) |
| $\beta_{LP^*}$ | 1.006 (0.995,1.018) | 1.095 (1.063,1.128) | 1.301 (1.217,1.391) | 1.930 (1.626,2.345) |
| | **MAR** | | | |
| MSE$_{apparent}$ | 2.993 (2.654,3.324) | 3.203 (2.812,3.596) | 4.508 (3.421,5.476) | 7.108 (5.313,8.166) |
| $\beta_{LP}$ | 1.026 (1.019,1.032) | 1.138 (1.094,1.172) | 1.409 (1.264,1.505) | 2.370 (1.765,3.066) |
| Tuning $\alpha$ | 0.761 (0.400,0.900) | 0.899 (0.900,0.900) | 0.900 (0.900,0.900) | 0.900 (0.900,0.900) |
| Tuning $\lambda$ | 0.035 (0.032,0.043) | 0.176 (0.147,0.210) | 0.529 (0.455,0.580) | 1.423 (1.199,1.528) |
| MSE$_{ext}$ | 3.173 (3.102,3.270) | 3.384 (3.208,3.606) | 4.932 (3.791,5.913) | 7.872 (5.919,8.876) |
| Optimism$_{ext}$ | -0.180 (-0.509,0.150) | -0.182 (-516,0.129) | -0.424 (-0.860,-0.017) | -0.765 (-1.307,-0.174) |
| Optimism$_{int}$ | -0.164 (-0.250,-0.096) | -0.122 (-0.192,-0.053) | -0.098 (-0.149,-0.043) | -0.056 (-0.113,-0.003) |
| MSE$_{corrected}$ | 3.158 (2.842,3.510) | 3.325 (2.948,3.734) | 4.606 (3.528,5.586) | 7.163 (5.382,8.198) |
| $\beta_{LP^*}$) | 1.010 (0.999,1.019) | 1.105 (1.073,1.134) | 1.327 (1.239,1.410) | 2.040 (1.684,2.446) |

Table A.5: **Accuracy** simulation study results for **MICE-Lasso** analysis with Harrell bootstrap validation: scenarios **S3** (assumption of moderation, without missing data) and **S4** (assumption of moderation, with missing data, complete outcome) based on 300 data sets of **20 variables** each (**n=1000**). Means of all estimates along with their corresponding 2.5th and 97.5th percentile values within parenthesis. The theoretical MSE is $1.74^2 = 3.028$

| **Estimates** | **Complete cases** | | | |
|---|---|---|---|---|
| | Best model | 1 SE tolerance | 3% tolerance | 15% tolerance |
| $\text{MSE}_{apparent}$ | 2.966 (2.697,3.251) | 3.009 (2.733,3.302) | 3.181 (2.885,3.500) | 3.911 (3.528,4.338) |
| $\beta_{LP}$ | 1.033 (1.026,1.040) | 1.052 (1.044,1.061) | 1.109 (1.097,1.125) | 1.240 (1.211,1.276) |
| Tuning $\lambda$ | 0.041 (0.031,0.049) | 0.066 (0.054,0.076) | 0.137 (0.119,0.149) | 0.298 (0.262,0.328) |
| $\text{MSE}_{ext}$ | 3.182 (3.114,3.268) | 3.217 (3.140,3.311) | 3.419 (3.292,3.555) | 4.312 (4.017,4.612) |
| $\text{Optimism}_{ext}$ | -0.216 (-0.487,0.092) | -0.208 (-0.475,0.101) | -0.238 (-0.541,0.063) | -0.401 (-0.765,-0.050) |
| $\text{Optimism}_{int}$ | -0.169 (-0.205,-0.138) | -0.143 (-0.177,-0.110) | -0.114 (-0.145,-0.084) | -0.114 (-0.148,-0.077) |
| $\text{MSE}_{corrected}$ | 3.135 (2.850,3.440) | 3.152 (2.863,3.457) | 3.295 (2.995,3.617) | 4.025 (3.638,4.472) |
| $\beta_{LP^*}$ | 1.019 (1.012,1.025) | 1.038 (1.029,1.046) | 1.097 (1.085,1.108) | 1.228 (1.200,1.259) |
| | **MCAR** | | | |
| $\text{MSE}_{apparent}$ | 3.320 (2.906,3.716) | 3.363 (2.941,3.769) | 3.558 (3.112,4.002) | 4.386 (3.836,4.927) |
| $\beta_{LP}$ | 1.030 (1.023,1.036) | 1.050 (1.043,1.058) | 1.110 (1.098,1.124) | 1.248 (1.217,1.284) |
| Tuning $\lambda$ | 0.039 (0.031,0.046) | 0.065 (0.055,0.073) | 0.141 (0.126,0.156) | 0.318 (0.284,0.351) |
| $\text{MSE}_{ext}$ | 3.338 (3.221,3.502) | 3.386 (3.252,3.570) | 3.652 (3.456,3.904) | 4.772 (4.335,5.275) |
| $\text{Optimism}_{ext}$ | -0.019 (-0.426,0.377) | -0.024 (-0.435,0.366) | -0.095 (-0.493,0.274) | -0.387 (-0.827,0.046) |
| $\text{Optimism}_{int}$ | -0.237 (-0.318,-0.148) | -0.209 (-0.287,-0.124) | -0.168 (-0.235,-0.091) | -0.154 (-0.207,-0.094) |
| $\text{MSE}_{corrected}$ | 3.557 (2.924,4.557) | 3.572 (2.971,4.629) | 3.726 (3.109,4.818) | 4.540 (3.910,6.027) |
| $\beta_{LP^*}$ | 1.013 (1.004,1.021) | 1.034 (1.024,1.043) | 1.094 (1.081,1.107) | 1.232 (1.202,1.264) |
| | **MAR** | | | |
| MSE apparent | 3.253 (2.899,3.611) | 3.296 (2.941,3.656) | 3.488 (3.114,3.873) | 4.297 (3.846,4.751) |
| $\beta_{LP}$ | 1.029 (1.023,1.035) | 1.050 (1.043,1.057) | 1.109 (1.096,1.122) | 1.245 (1.216,1.277) |
| Tuning $\lambda$ | 0.038 (0.030,0.045) | 0.063 (0.054,0.073) | 0.138 (0.124,0.151) | 0.311 (0.287,0.337) |
| $\text{MSE}_{ext}$ | 3.310 (3.193,3.446) | 3.350 (3.217,3.488) | 3.594 (3.384,3.801) | 4.659 (4.258,5.067) |
| $\text{Optimism}_{ext}$ | -0.057 (-0.407,0.295) | -0.054 (-0.405,0.289) | -0.106 (-0.482,0.234) | -0.362 (-0.787,0.081) |
| $\text{Optimism}_{int}$ | -0.224 (-0.304,-0.147) | -0.197 (-0.276,-0.125) | -0.159 (-0.235,-0.090) | -0.150 (-0.210,-0.087) |
| $\text{MSE}_{corrected}$ | 3.610 (2.886,4.346) | 3.670 (2.929,4.429) | 3.847 (3.088,4.641) | 4.837 (3.938,5.964) |
| $\beta_{LP^*}$ | 1.014 (1.004,1.023) | 1.034 (1.024,1.044) | 1.094 (1.080,1.108) | 1.230 (1.202,1.260) |

Table A.6: **Accuracy** simulation study results for **MICE-Elasticnet** analysis with Harrell bootstrap validation: scenarios **S3** (assumption of moderation, without missing data) and **S4** (assumption of moderation, with missing data, complete outcome) based on 300 data sets of **20 variables** each (**n=1000**). Means of all estimates along with their corresponding 2.5th and 97.5th percentile values within parenthesis. The theoretical MSE is $1.74^2 = 3.028$

| **Estimates** | **Complete cases** | | | |
|---|---|---|---|---|
| | Best model | 1 SE tolerance | 3% tolerance | 15% tolerance |
| $MSE_{apparent}$ | 2.966 (2.697,3.254) | 3.186 (2.849,3.550) | 5.410 (4.367,5.953) | 9.898 (8.468,10.788) |
| $\beta_{LP}$ | 1.033 (1.027,1.040) | 1.111 (1.086,1.157) | 1.398 (1.317,1.468) | 2.228 (1.793,2.548) |
| Tuning $\alpha$ | 0.881 (0.800,0.900) | 0.900 (0.900,0.900) | 0.900 (0.900,0.900) | 0.900 (0.900,0.900) |
| Tuning $\lambda$ | 0.046 (0.040,0.051) | 0.154 (0.135,0.220) | 0.565 (0.455,0.580) | 1.497 (1.199,1.528) |
| $MSE_{ext}$ | 3.183 (3.114,3.267) | 3.426 (3.260,3.685) | 6.119 (4.858,6.945) | 11.203 (9.975,12.007) |
| $Optimism_{ext}$ | -0.217 (-0.489,0.091) | -0.240 (-0.534,0.057) | -0.709 (-1.242,-0.238) | -1.305 (-2.137,-0.498) |
| $Optimism_{int}$ | -0.171 (-0.205,-0.140) | -0.121 (-0.153,-0.088) | -0.127 (-0.171,-0.089) | -0.078 (-0.158,-0.002) |
| $MSE_{corrected}$ | 3.137 (2.855,3.442) | 3.307 (2.952,3.691) | 5.537 (4.502,6.079) | 9.976 (8.547,10.838) |
| $beta_{LP*}$ | 1.019 (1.014,1.025) | 1.102 (1.076,1.121) | 1.323 (1.230,1.397) | 1.942 (1.628,2.238) |
| | **MCAR** | | | |
| $MSE_{apparent}$ | 3.319 (2.906,3.716) | 3.652 (3.168,4.123) | 5.806 (4.781,6.742) | 9.981 (8.221,11.591) |
| $\beta_{LP}$ | 1.030 (1.024,1.036) | 1.134 (1.105,1.170) | 1.405 (1.325,1.498) | 2.356 (1.895,3.179) |
| Tuning $\alpha$ | 0.841 (0.619,0.900) | 0.900 (0.900,0.900) | 0.900 (0.900,0.900) | 0.900 (0.900,0.900) |
| Tuning $\lambda$ | 0.047 (0.037,0.060) | 0.198 (0.150,0.259) | 0.618 (0.542,0.739) | 1.605 (1.430,1.906) |
| $MSE_{ext}$ | 3.339 (3.220,3.504) | 3.787 (3.501,4.147) | 6.547 (5.188,7.451) | 11.298 (9.376,12.730) |
| $Optimism_{ext}$ | -0.020 (-0.426,0.378) | -0.135 (-0.522,0.250) | -0.741 (-1.290,-0.160) | -1.316 (-2.206,-0.482) |
| $Optimism_{int}$ | -0.239 (-0.319,-0.149) | -0.175 (-0.247,-0.096) | -0.153 (-0.212,-0.092) | -0.092 (-0.169,-0.015) |
| $MSE_{corrected}$ | 3.558 (3.104,4.002) | 3.827 (3.316,4.351) | 5.959 (4.930,6.911) | 10.073 (8.334,11.678) |
| $\beta_{LP^*}$ | 1.015 (1.006,1.023) | 1.106 (1.080,1.129) | 1.318 (1.239,1.394) | 1.991 (1.700,2.390) |
| | **MAR** | | | |
| $MSE_{apparent}$ | 3.253 (2.902,3.610) | 3.584 (3.143,4.046) | 5.666 (4.671,6.591) | 9.781 (8.063,11.094) |
| $\beta_{LP}$ | 1.030 (1.024,1.035) | 1.133 (1.104,1.167) | 1.395 (1.321,1.461) | 2.264 (1.905,2.843) |
| Tuning $\alpha$ | 0.831 (0.609,0.900) | 0.900 (0.900,0.900) | 0.900 (0.900,0.900) | 0.900 (0.900,0.900) |
| Tuning $\lambda$ | 0.046 (0.037,0.063) | 0.197 (0.149,0.258) | 0.603 (0.536,0.715) | 1.568 (1.445,1.822) |
| $MSE_{ext}$ | 3.311 (3.194,3.446) | 3.725 (3.484,4.040) | 6.373 (5.163,7.289) | 11.072 (9.293,12.179) |
| $Optimism_{ext}$ | -0.059 (-0.407,0.294) | -0.141 (-0.503,0.203) | -0.707 (-1.262,-0.157) | -1.292 (-2.202,-0.489) |
| $Optimism_{int}$ | -0.225 (-0.306,-0.149) | -0.166 (-0.240,-0.098) | -0.150 (-0.206,-0.097) | -0.093 (-0.165,-0.014) |
| $MSE_{corrected}$ | 3.478 (3.088,3.886) | 3.750 (3.262,4.212) | 5.816 (4.842,6.754) | 9.874 (8.116,11.142) |
| $beta_{LP*}$ | 1.016 (1.006,1.025) | 1.106 (1.081,1.127) | 1.312 (1.233,1.385) | 1.945 (1.657,2.263) |

Table A.7: **Accuracy** simulation study results for **MICE-Lasso** analysis with Harrell bootstrap validation: scenarios **S3** (assumption of moderation, without missing data) and **S5** (assumption of moderation, with missing data also in the outcome) based on 300 data sets of **20 variables** each (**n=1000**). Means of all estimates along with their corresponding 2.5th and 97.5th percentile values within parenthesis. The theoretical MSE is $1.74^2 = 3.028$

| **Estimates** | **Complete cases** | | | |
|---|---|---|---|---|
| | Best model | 1 SE tolerance | 3% tolerance | 15% tolerance |
| $MSE_{apparent}$ | 2.966 (2.697,3.251) | 3.009 (2.733,3.302) | 3.181 (2.885,3.500) | 3.911 (3.528,4.338) |
| $\beta_{LP}$ | 1.033 (1.026,1.040) | 1.052 (1.044,1.061) | 1.109 (1.097,1.125) | 1.240 (1.211,1.276) |
| Tuning $\lambda$ | 0.041 (0.031,0.049) | 0.066 (0.054,0.076) | 0.137 (0.119,0.149) | 0.298 (0.262,0.328) |
| $MSE_{ext}$ | 3.182 (3.114,3.268) | 3.217 (3.140,3.311) | 3.419 (3.292,3.555) | 4.312 (4.017,4.612) |
| $Optimism_{ext}$ | -0.216 (-0.487,0.092) | -0.208 (-0.475,0.101) | -0.238 (-0.541,0.063) | -0.401 (-0.765,-0.050) |
| $Optimism_{int}$ | -0.169 (-0.205,-0.138) | -0.143 (-0.177,-0.110) | -0.114 (-0.145,-0.084) | -0.114 (-0.148,-0.077) |
| $MSE_{corrected}$ | 3.135 (2.850,3.440) | 3.152 (2.863,3.457) | 3.295 (2.995,3.617) | 4.025 (3.638,4.472) |
| $\beta_{LP^*}$ | 1.019 (1.012,1.025) | 1.038 (1.029,1.046) | 1.097 (1.085,1.108) | 1.228 (1.200,1.259) |
| | **MCAR** | | | |
| $MSE_{apparent}$ | 3.514 (3.022,3.940) | 3.558 (3.059,3.989) | 3.760 (3.237,4.216) | 4.635 (3.965,5.219) |
| $\beta_{LP}$ | 1.029 (1.023,1.035) | 1.050 (1.043,1.059) | 1.112 (1.098,1.128) | 1.257 (1.217,1.302) |
| Tuning $\lambda$ | 0.038 (0.031,0.046) | 0.065 (0.055,0.074) | 0.142 (0.125,0.158) | 0.326 (0.290,0.365) |
| $MSE_{ext}$ | 3.451 (3.301,3.649) | 3.499 (3.333,3.720) | 3.771 (3.540,4.076) | 4.949 (4.466,5.568) |
| $Optimism_{ext}$ | 0.063 (-0.456,0.510) | 0.059 (-0.440,0.500) | -0.011 (-0.531,0.427) | -0.314 (-0.832,0.156) |
| $Optimism_{int}$ | -0.299 (-0.423,-0.192) | -0.269 (-0.386,-0.167) | -0.220 (-0.331,-0.124) | -0.191 (-0.292,-0.108) |
| $MSE_{corrected}$ | 3.813 (3.232,4.280) | 3.827 (3.243,4.288) | 3.980 (3.382,4.460) | 4.826 (4.110,5.390) |
| $\beta_{LP^*}$ | 1.011 (0.999,1.021) | 1.032 (1.019,1.043) | 1.093 (1.079,1.110) | 1.236 (1.206,1.274) |
| | **MAR** | | | |
| $MSE_{apparent}$ | 3.452 (3.046,3.855) | 3.496 (3.082,3.907) | 3.697 (3.259,4.124) | 4.555 (4.003,5.115) |
| $\beta_{LP}$ | 1.027 (1.022,1.033) | 1.048 (1.040,1.057) | 1.109 (1.096,1.124) | 1.249 (1.216,1.282) |
| Tuning $\lambda$ | 0.035 (0.029,0.043) | 0.061 (0.052,0.071) | 0.137 (0.122,0.152) | 0.315 (0.287,0.347) |
| $MSE_{ext}$ | 3.448 (3.309,3.638) | 3.473 (3.321,3.645) | 3.685 (3.486,3.887) | 4.740 (4.299,5.248) |
| $Optimism_{ext}$ | 0.004 (-0.473,0.445) | 0.022 (-0.437,0.443) | 0.012 (-0.430,0.421) | -0.184 (-0.670,0.341) |
| $Optimism_{int}$ | -0.277 (-0.387,-0.171) | -0.251 (-0.355,-0.148) | -0.209 (-0.307,-0.107) | -0.188 (-0.298,-0.089) |
| $MSE_{corrected}$ | 3.729 (3.245,4.196) | 3.746 (3.267,4.211) | 3.906 (3.414,4.388) | 4.743 (4.186,5.335) |
| $\beta_{LP^*}$ | 1.012 (1.001,1.023) | 1.033 (1.021,1.045) | 1.093 (1.078,1.109) | 1.232 (1.202,1.266) |

Table A.8: **Accuracy** simulation study results for **MICE-Elasticnet** analysis with Harrell bootstrap validation: scenarios **S3** (assumption of moderation, without missing data) and **S5** (assumption of moderation, with missing data also in the outcome) based on 300 data sets of **20 variables** each (**n=1000**). Means of all estimates along with their corresponding 2.5th and 97.5th percentile values within parenthesis. The theoretical MSE is $1.74^2 = 3.028$

| **Estimates** | **Complete cases** | | | |
|---|---|---|---|---|
| | Best model | 1 SE tolerance | 3% tolerance | 15% tolerance |
| $MSE_{apparent}$ | 2.966 (2.697,3.254) | 3.186 (2.849,3.550) | 5.410 (4.367,5.953) | 9.898 (8.468,10.788) |
| $\beta_{LP}$ | 1.033 (1.027,1.040) | 1.111 (1.086,1.157) | 1.398 (1.317,1.468) | 2.228 (1.793,2.548) |
| Tuning $\alpha$ | 0.881 (0.800,0.900) | 0.900 (0.900,0.900) | 0.900 (0.900,0.900) | 0.900 (0.900,0.900) |
| Tuning $\lambda$ | 0.046 (0.040,0.051) | 0.154 (0.135,0.220) | 0.565 (0.455,0.580) | 1.497 (1.199,1.528) |
| $MSE_{ext}$ | 3.183 (3.114,3.267) | 3.426 (3.260,3.685) | 6.119 (4.858,6.945) | 11.203 (9.975,12.007) |
| $Optimism_{ext}$ | -0.217 (-0.489,0.091) | -0.240 (-0.534,0.057) | -0.709 (-1.242,-0.238) | -1.305 (-2.137,-0.498) |
| $Optimism_{int}$ | -0.171 (-0.205,-0.140) | -0.121 (-0.153,-0.088) | -0.127 (-0.171,-0.089) | -0.078 (-0.158,-0.002) |
| $MSE_{corrected}$ | 3.137 (2.855,3.442) | 3.307 (2.952,3.691) | 5.537 (4.502,6.079) | 9.976 (8.547,10.838) |
| $beta_{LP*}$ | 1.019 (1.014,1.025) | 1.102 (1.076,1.121) | 1.323 (1.230,1.397) | 1.942 (1.628,2.238) |
| | **MCAR** | | | |
| $MSE_{apparent}$ | 3.514 (3.022,3.941) | 3.873 (3.337,4.412) | 5.915 (4.984,6.974) | 9.991 (8.354,11.560) |
| $\beta_{LP}$ | 1.030 (1.024,1.036) | 1.141 (1.112,1.172) | 1.407 (1.326,1.504) | 2.374 (1.918,3.231) |
| Tuning $\alpha$ | 0.807 (0.570,0.890) | 0.900 (0.900,0.900) | 0.900 (0.900,0.900) | 0.900 (0.900,0.900) |
| Tuning $\lambda$ | 0.049 (0.037,0.072) | 0.214 (0.159,0.282) | 0.631 (0.542,0.739) | 1.655 (1.463,1.947) |
| $MSE_{ext}$ | 3.451 (3.301,3.651) | 3.929 (3.593,4.307) | 6.543 (5.368,7.638) | 11.266 (9.496,12.823) |
| $Optimism_{ext}$ | 0.063 (-0.456,0.507) | -0.056 (-0.574,0.377) | -0.628 (-1.253,0.009) | -1.275 (-2.336,-0.344) |
| $Optimism_{int}$ | -0.300 (-0.425,-0.193) | -0.224 (-0.338,-0.127) | -0.184 (-0.285,-0.092) | -0.104 (-0.261,0.034) |
| $MSE_{corrected}$ | 3.814 (3.232,4.278) | 4.097 (3.496,4.664) | 6.099 (5.134,7.173) | 10.095 (8.474,11.754) |
| $\beta_{LP^*}$ | 1.013 (1.001,1.024) | 1.107 (1.083,1.131) | 1.320 (1.251,1.392) | 2.001 (1.720,2.391) |
| | **MAR** | | | |
| $MSE_{apparent}$ | 3.452 (3.047,3.854) | 3.805 (3.332,4.296) | 5.620 (4.502,6.699) | 9.568 (7.046,11.260) |
| $\beta_{LP}$ | 1.029 (1.023,1.035) | 1.137 (1.103,1.170) | 1.381 (1.264,1.470) | 2.156 (1.689,2.715) |
| Tuning $\alpha$ | 0.751 (0.420,0.885) | 0.900 (0.900,0.900) | 0.900 (0.900,0.900) | 0.900 (0.900,0.900) |
| Tuning $\lambda$ | 0.050 (0.037,0.083) | 0.220 (0.168,0.277) | 0.615 (0.542,0.723) | 1.605 (1.463,1.906) |
| $MSE_{ext}$ | 3.447 (3.312,3.635) | 3.816 (3.549,4.139) | 6.021 (4.450,7.183) | 10.518 (7.642,12.106) |
| $Optimism_{ext}$ | 0.005 (-0.470,0.444) | -0.011 (-0.451,0.405) | -0.401 (-1.059,0.225) | -0.951 (-2.110,0.107) |
| $Optimism_{int}$ | -0.278 (-0.387,-0.173) | -0.215 (-0.312,-0.111) | -0.183 (-0.292,-0.081) | -0.114 (-0.271,0.030) |
| $MSE_{corrected}$ | 3.730 (3.247,4.196) | 4.020 (3.498,4.550) | 5.804 (4.673,6.867) | 9.681 (7.269,11.371) |
| $beta_{LP*}$ | 1.014 (1.004,1.025) | 1.103 (1.079,1.127) | 1.299 (1.223,1.370) | 1.882 (1.631,2.185) |

Table A.9: **Accuracy** simulation study results for **MICE-Lasso** analysis with Harrell bootstrap validation: scenarios S3 (assumption of moderation, without missing data) and **S6** (assumption of moderation, missing data, interaction terms in the imputation model), based on 300 data sets of **20 variables** each (**n=1000**). Means of all estimates along with their corresponding 2.5th and 97.5th percentile values within parenthesis. The theoretical MSE is $1.74^2 = 3.028$

| **Estimates** | **Complete cases** | | | |
|---|---|---|---|---|
| | Best model | 1 SE tolerance | 3% tolerance | 15% tolerance |
| $\text{MSE}_{apparent}$ | 2.966 (2.697,3.251) | 3.009 (2.733,3.302) | 3.181 (2.885,3.500) | 3.911 (3.528,4.338) |
| $\beta_{LP}$ | 1.033 (1.026,1.040) | 1.052 (1.044,1.061) | 1.109 (1.097,1.125) | 1.240 (1.211,1.276) |
| Tuning $\lambda$ | 0.041 (0.031,0.049) | 0.066 (0.054,0.076) | 0.137 (0.119,0.149) | 0.298 (0.262,0.328) |
| $\text{MSE}_{ext}$ | 3.182 (3.114,3.268) | 3.217 (3.140,3.311) | 3.419 (3.292,3.555) | 4.312 (4.017,4.612) |
| $\text{Optimism}_{ext}$ | -0.216 (-0.487,0.092) | -0.208 (-0.475,0.101) | -0.238 (-0.541,0.063) | -0.401 (-0.765,-0.050) |
| $\text{Optimism}_{int}$ | -0.169 (-0.205,-0.138) | -0.143 (-0.177,-0.110) | -0.114 (-0.145,-0.084) | -0.114 (-0.148,-0.077) |
| $\text{MSE}_{corrected}$ | 3.135 (2.850,3.440) | 3.152 (2.863,3.457) | 3.295 (2.995,3.617) | 4.025 (3.638,4.472) |
| $\beta_{LP^*}$ | 1.019 (1.012,1.025) | 1.038 (1.029,1.046) | 1.097 (1.085,1.108) | 1.228 (1.200,1.259) |
| | **MCAR** | | | |
| $\text{MSE}_{apparent}$ | 3.256 (2.832,3.721) | 3.306 (2.876,3.771) | 3.508 (3.060,3.988) | 4.315 (3.759,4.932) |
| $\beta_{LP}$ | 1.021 (1.013,1.030) | 1.040 (1.025,1.052) | 1.101 (1.081,1.119) | 1.240 (1.207,1.276) |
| Tuning $\lambda$ | 0.028 (0.017,0.039) | 0.052 (0.032,0.068) | 0.129 (0.102,0.149) | 0.305 (0.270,0.340) |
| $\text{MSE}_{ext}$ | 3.301 (3.188,3.430) | 3.319 (3.198,3.467) | 3.546 (3.341,3.792) | 4.619 (4.139,5.152) |
| $\text{Optimism}_{ext}$ | -0.044 (-0.455,0.409) | -0.013 (-0.418,0.416) | -0.038 (-0.412,0.354) | -0.304 (-0.736,0.102) |
| $\text{Optimism}_{int}$ | -0.296 (-0.400,-0.202) | -0.254 (-0.350,-0.159) | -0.171 (-0.247,-0.093) | -0.144 (-0.206,-0.081) |
| $\text{MSE}_{corrected}$ | 3.552 (3.120,4.076) | 3.560 (3.122,4.088) | 3.678 (3.223,4.201) | 4.460 (3.887,5.101) |
| $\beta_{LP^*}$ | 1.000 (0.990,1.009) | 1.017 (1.005,1.029) | 1.077 (1.057,1.096) | 1.218 (1.185,1.252) |
| | **MAR** | | | |
| $\text{MSE}_{apparent}$ | 3.146 (2.784,3.532) | 3.194 (2.826,3.588) | 3.387 (3.005,3.805) | 4.163 (3.694,4.680) |
| $\beta_{LP}$ | 1.022 (1.015,1.029) | 1.042 (1.030,1.052) | 1.101 (1.086,1.115) | 1.236 (1.207,1.270) |
| Tuning $\lambda$ | 0.029 (0.019,0.039) | 0.053 (0.040,0.065) | 0.128 (0.111,0.145) | 0.298 (0.271,0.328) |
| $\text{MSE}_{ext}$ | 3.269 (3.168,3.412) | 3.288 (3.181,3.433) | 3.499 (3.315,3.709) | 4.495 (4.070,4.924) |
| $\text{Optimism}_{ext}$ | -0.122 (-0.453,0.288) | -0.094 (-0.418,0.289) | -0.113 (-0.446,0.267) | -0.331 (-0.720,0.051) |
| $\text{Optimism}_{int}$ | -0.285 (-0.377,-0.203) | -0.244 (-0.329,-0.163) | -0.169 (-0.245,-0.095) | -0.145 (-0.208,-0.075) |
| $\text{MSE}_{corrected}$ | 3.432 (3.032,3.878) | 3.439 (3.046,3.888) | 3.556 (3.151,4.026) | 4.308 (3.833,4.842) |
| $\beta_{LP^*}$ | 1.003 (0.994,1.013) | 1.020 (1.009,1.032) | 1.079 (1.064,1.094) | 1.216 (1.191,1.246) |

Figure A.1: Comparison of **variable inclusion frequency** by **MICE-Lasso** (ML) run on 300 simulated **20-covariate** datasets with 250 observations for the scenarios with moderation assumption **S3** (without missing data), **S4** (with missing data, complete outcome), **S5** ( with missing data also in the outcome) and **S6** (missing data, complete outcome and interaction terms in the imputation model) with **MAR** data. ML variable inclusion frequencies are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning.

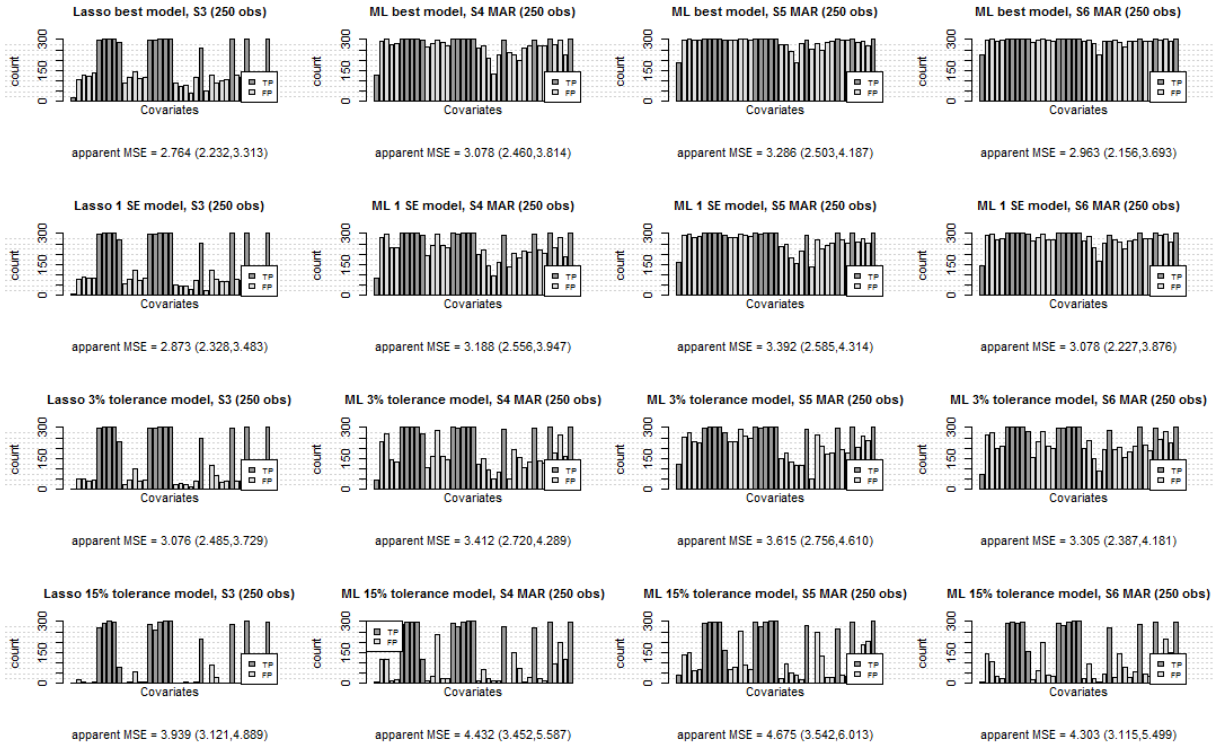


Figure A.2: Comparison of **variable inclusion frequency** by **MICE-Lasso** (ML) run on 300 simulated **20-covariate** datasets with 1000 observations for the scenarios with moderation assumption **S3** (without missing data), **S4** (with missing data, complete outcome), **S5** ( with missing data also in the outcome) and **S6** (missing data, complete outcome and interaction terms in the imputation model) with **MCAR** data. ML variable inclusion frequencies are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning.

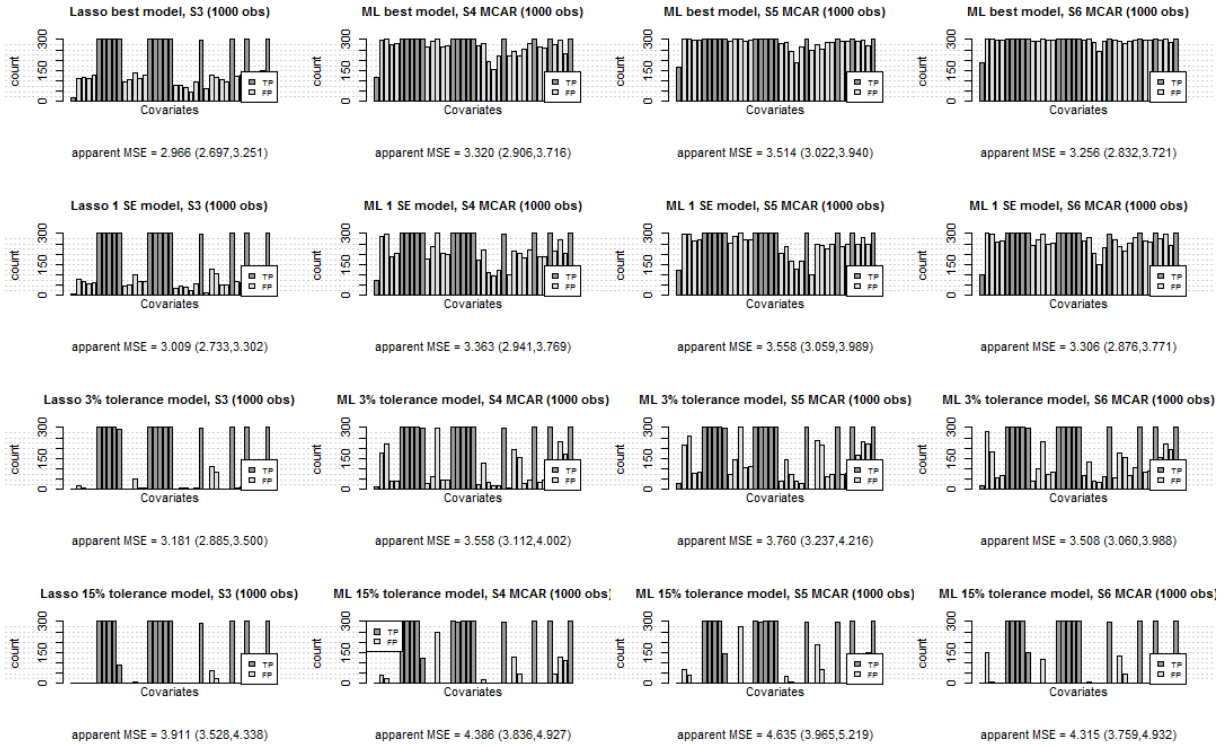


Figure A.3: Comparison of **variable inclusion frequency** by **MICE-Lasso** (ML) run on 300 simulated **20-covariate** datasets with 1000 observations for the scenarios with moderation assumption **S3** (without missing data), **S4** (with missing data, complete outcome), **S5** ( with missing data also in the outcome) and **S6** (missing data, complete outcome and interaction terms in the imputation model) with **MAR** data. ML variable inclusion frequencies are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning.

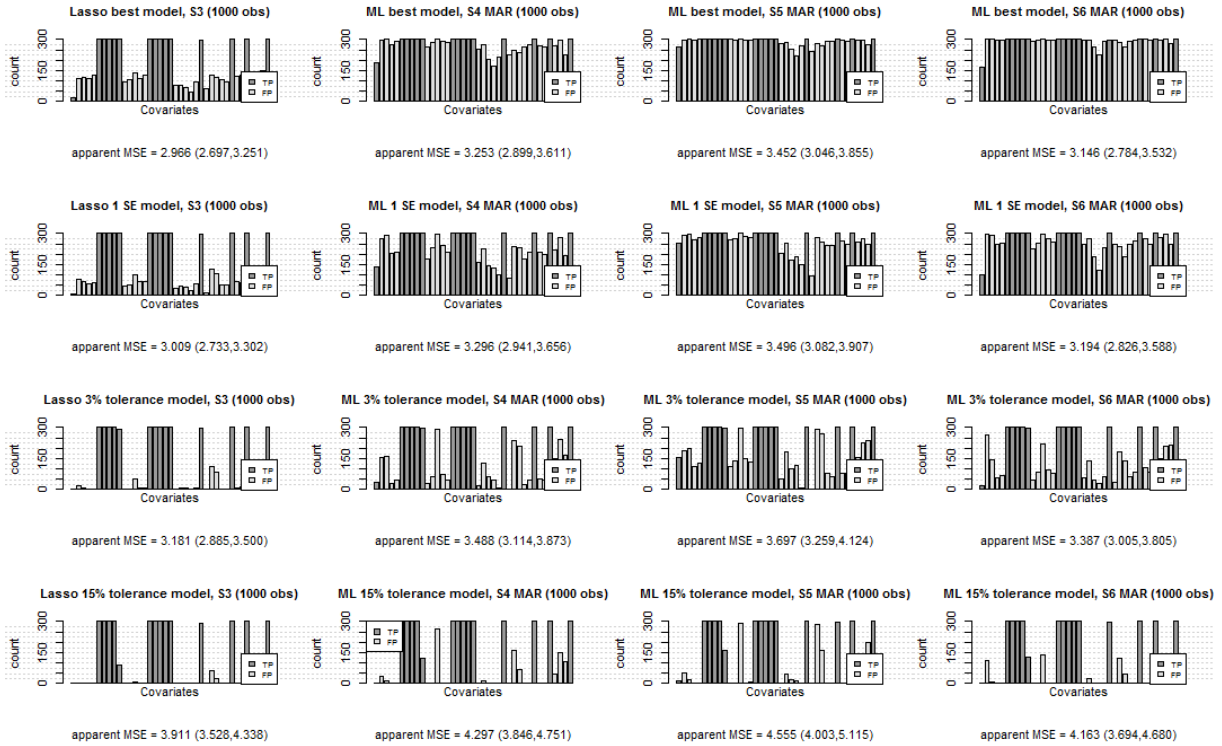


Table A.10: **Accuracy** simulation study results for **MICE-Lasso** analysis with Harrell (1996) bootstrap validation: scenarios **S3** (assumption of moderation, complete data) and **S5** (assumption of moderation, missing data also in the outcome), based on 300 data sets of **100 variables** each (**n=500**) with between-covariate **correlation of 0.8**. Means of all estimates along with their corresponding 2.5th and 97.5th percentile values within parenthesis. The theoretical MSE is 1.

| **Estimates** | **Complete cases** | | | |
|---|---|---|---|---|
| | Best model | 1 SE tolerance | 3% tolerance | 15% tolerance |
| $MSE_{apparent}$ | 1.000 (0.852,1.146) | 1.053 (0.916,1.215) | 1.322 (1.153,1.501) | 1.528 (1.337,1.738) |
| $\beta_{LP}$ | 1.054 (1.044,1.065) | 1.063 (1.050,1.077) | 1.127 (1.099,1.155) | 1.188 (1.145,1.242) |
| Tuning $\lambda$ | 0.044 (0.031,0.054) | 0.141 (0.107,0.187) | 0.056 (0.039,0.068) | 0.224 (0.167,0.293) |
| $MSE_{ext}$ | 1.112 (1.041,1.183) | 1.128 (1.054,1.212) | 1.320 (1.188,1.465) | 1.516 (1.336,1.718) |
| $Optimism_{ext}$ | -0.112 (-0.255,0.044) | -0.075 (-0.206,0.075) | 0.002 (-0.159,0.160) | 0.012 (-0.179,0.203) |
| $Optimism_{int}$ | -0.302 (-0.351,-0.259) | -0.255 (-0.303,-0.211) | -0.122 (-0.160,-0.090) | -0.090 (-0.120,-0.063) |
| $MSE_{corrected}$ | 1.302 (1.134,1.491) | 1.308 (1.152,1.494) | 1.444 (1.263,1.629) | 1.617 (1.414,1.832) |
| $\beta_{LP^*}$ | 1.016 (1.010,1.024) | 1.026 (1.019,1.036) | 1.078 (1.061,1.099) | 1.122 (1.096,1.156) |
| | | **MCAR** | | |
| $MSE_{apparent}$ | 1.168 (0.977,1.382) | 1.176 (0.981,1.389) | 1.368 (1.133,1.617) | 1.586 (1.309,1.892) |
| $\beta_{LP}$ | 1.014 (1.000,1.026) | 1.018 (1.004,1.033) | 1.068 (1.049,1.092) | 1.107 (1.077,1.142) |
| Tuning $\lambda$ | 0.016 (0.016,0.018) | 0.054 (0.037,0.078) | 0.018 (0.016,0.022) | 0.106 (0.071,0.153) |
| $MSE_{ext}$ | 1.542 (1.363,1.745) | 1.526 (1.356,1.727) | 1.429 (1.276,1.597) | 1.507 (1.333,1.759) |
| $Optimism_{ext}$ | -0.374 (-0.629,-0.122) | -0.350 (-0.608,-0.090) | -0.061 (-0.316,0.211) | 0.078 (-0.197,0.382) |
| $Optimism_{int}$ | -0.368 (-0.438,-0.306) | -0.307 (-0.378,-0.246) | -0.143 (-0.189,-0.098) | -0.103 (-0.144,-0.065) |
| $MSE_{corrected}$ | 1.536 (1.310,1.768) | 1.483 (1.261,1.716) | 1.510 (1.274,1.763) | 1.689 (1.413,1.997) |
| $\beta_{LP^*}$ | 1.004 (0.993,1.015) | 1.016 (1.005,1.028) | 1.078 (1.059,1.103) | 1.142 (1.105,1.194) |
| | | **MAR** | | |
| $MSE_{apparent}$ | 1.091 (0.928,1.276) | 1.102 (0.940,1.291) | 1.325 (1.120,1.569) | 1.551 (1.306,1.832) |
| $\beta_{LP}$ | 1.017 (1.004,1.030) | 1.022 (1.006,1.035) | 1.074 (1.054,1.098) | 1.116 (1.087,1.157) |
| Tuning $\lambda$ | 0.016 (0.016,0.019) | 0.061 (0.043,0.089) | 0.019 (0.016,0.024) | 0.117 (0.078,0.176) |
| $MSE_{ext}$ | 1.522 (1.363,1.717) | 1.502 (1.351,1.686) | 1.425 (1.285,1.608) | 1.530 (1.344,1.780) |
| $Optimism_{ext}$ | -0.431 (-0.651,-0.222) | -0.400 (-0.616,-0.185) | -0.099 (-0.321,0.138) | 0.021 (-0.221,0.270) |
| $Optimism_{int}$ | -0.529 (-0.646,-0.417) | -0.526 (-0.642,-0.414) | -0.398 (-0.489,-0.310) | -0.314 (-0.393,-0.234) |
| $MSE_{corrected}$ | 1.620 (1.350,1.911) | 1.628 (1.357,1.923) | 1.724 (1.456,2.040) | 1.865 (1.575,2.191) |
| $\beta_{LP^*}$ | 0.979 (0.956,0.996) | 0.980 (0.957,0.997) | 1.024 (1.005,1.041) | 1.055 (1.036,1.074) |

Table A.11: **Accuracy** simulation study results for **MICE-Elasticnet** analysis with Harrell (1996) bootstrap validation: scenarios **S3** (assumption of moderation, complete data) and **S5** (assumption of moderation, missing data also in the outcome), based on 300 data sets of **100 variables** each (**n=500**) with between-covariate **correlation of 0.8**. Means of all estimates along with their corresponding 2.5th and 97.5th percentile values within parenthesis. The theoretical MSE is $1$.

| **Estimates** | **Complete cases** | | | |
|---|---|---|---|---|
| | Best model | 1 SE tolerance | 3% tolerance | 15% tolerance |
| $MSE_{apparent}$ | 0.998 (0.842,1.148) | 1.092 (0.935,1.262) | 2.134 (1.806,2.549) | 2.944 (2.384,3.617) |
| $\beta_{LP}$ | 1.054 (1.044,1.066) | 1.073 (1.055,1.095) | 1.435 (1.274,1.677) | 2.140 (1.595,3.105) |
| Tuning $\alpha$ | 0.897 (0.800,0.900) | 0.900 (0.900,0.900) | 0.900 (0.900,0.900) | 0.900 (0.900,0.900) |
| Tuning $\lambda$ | 0.048 (0.032,0.065) | 0.546 (0.357,0.739) | 0.075 (0.051,0.106) | 0.976 (0.739,1.199) |
| $MSE_{ext}$ | 1.113 (1.040,1.184) | 1.150 (1.062,1.261) | 2.151 (1.733,2.541) | 3.018 (2.369,3.888) |
| $Optimism_{ext}$ | -0.115 (-0.258,0.049) | -0.058 (-0.193,0.102) | -0.017 (-0.285,0.264) | -0.074 (-0.468,0.318) |
| $Optimism_{int}$ | -0.313 (-0.373,-0.261) | -0.209 (-0.264,-0.159) | -0.092 (-0.137,-0.060) | -0.077 (-0.114,-0.046) |
| $MSE_{corrected}$ | 1.311 (1.134,1.491) | 1.302 (1.152,1.494) | 2.226 (1.263,1.629) | 3.021 (1.414,1.832) |
| $\beta_{LP^*}$ | 1.018 (1.011,1.026) | 1.047 (1.033,1.066) | 1.202 (1.130,1.304) | 1.379 (1.223,1.647) |
| **MCAR** | | | | |
| $MSE_{apparent}$ | 1.156 (0.963,1.369) | 1.165 (0.967,1.383) | 1.362 (1.087,1.666) | 1.582 (1.246,1.963) |
| $\beta_{LP}$ | 0.992 (0.972,1.014) | 1.011 (0.985,1.043) | 1.086 (1.048,1.140) | 1.138 (1.081,1.231) |
| Tuning $\alpha$ | 0.172 (0.115,0.270) | 0.819 (0.670,0.900) | 0.290 (0.165,0.465) | 0.894 (0.855,0.900) |
| Tuning $\lambda$ | 0.040 (0.032,0.059) | 0.261 (0.151,0.398) | 0.062 (0.041,0.092) | 0.525 (0.298,0.814) |
| $MSE_{ext}$ | 1.686 (1.459,1.924) | 1.609 (1.405,1.850) | 1.501 (1.351,1.678) | 1.603 (1.402,1.885) |
| $Optimism_{ext}$ | -0.530 (-0.831,-0.232) | -0.444 (-0.755,-0.136) | -0.139 (-0.420,0.159) | -0.021 (-0.285,0.284) |
| $Optimism_{int}$ | -0.407 (-0.482,-0.339) | -0.314 (-0.394,-0.228) | -0.159 (-0.219,-0.106) | -0.115 (-0.162,-0.069) |
| $MSE_{corrected}$ | 1.564 (1.310,1.768) | 1.478 (1.261,1.716) | 1.521 (1.274,1.763) | 1.698 (1.413,1.997) |
| $\beta_{LP^*}$ | 1.005 (0.994,1.016) | 1.029 (1.013,1.049) | 1.128 (1.069,1.232) | 1.862 (1.121,2.481) |
| **MAR** | | | | |
| $MSE_{apparent}$ | 1.074 (0.906,1.263) | 1.090 (0.920,1.287) | 1.331 (1.083,1.637) | 1.575 (1.258,1.955) |
| $\beta_{LP}$ | 0.998 (0.975,1.017) | 1.020 (0.992,1.048) | 1.101 (1.058,1.166) | 1.162 (1.094,1.295) |
| Tuning $\alpha$ | 0.191 (0.120,0.310) | 0.844 (0.730,0.900) | 0.327 (0.200,0.505) | 0.897 (0.870,0.900) |
| Tuning $\lambda$ | 0.041 (0.032,0.063) | 0.291 (0.176,0.449) | 0.067 (0.044,0.102) | 0.586 (0.348,0.928) |
| $MSE_{ext}$ | 1.663 (1.456,1.944) | 1.576 (1.399,1.825) | 1.505 (1.343,1.702) | 1.647 (1.428,1.954) |
| $Optimism_{ext}$ | -0.588 (-0.884,-0.327) | -0.486 (-0.789,-0.217) | -0.175 (-0.450,0.092) | -0.072 (-0.357,0.194) |
| $Optimism_{int}$ | -0.671 (-0.809,-0.549) | -0.639 (-0.764,-0.523) | -0.475 (-0.569,-0.391) | -0.390 (-0.473,-0.312) |
| $MSE_{corrected}$ | 1.745 (1.472,2.066) | 1.729 (1.452,2.057) | 1.805 (1.518,2.142) | 1.965 (1.640,2.365) |
| $\beta_{LP^*}$ | 0.935 (0.908,0.955) | 0.947 (0.921,0.968) | 1.014 (0.994,1.032) | 1.051 (1.032,1.071) |

## A.2 MissForest-Lasso and MissForest-Elasticnet simulation results

Table A.12: **Accuracy** simulation study results for **MissForest-Lasso** analysis with Harrell bootstrap validation: scenarios **S1** (without missing data, no assumption of moderation) and **S2** (with missing data, complete outcome, no assumption of moderation) based on 300 data sets of **20 variables** each (**n=1000**). Means of all estimates along with their corresponding 2.5th and 97.5th percentile values within parenthesis. The theoretical MSE is $1.74^2 = 3.028$

| **Estimates** | **Complete cases** | | | |
|---|---|---|---|---|
| | Best model | 1 SE tolerance | 3% tolerance | 15% tolerance |
| $MSE_{apparent}$ | 2.985 (2.718,3.274) | 3.024 (2.749,3.317) | 3.189 (2.887,3.509) | 3.922 (3.542,4.295) |
| $\beta_{LP}$ | 1.032 (1.022,1.041) | 1.060 (1.050,1.072) | 1.137 (1.118,1.156) | 1.309 (1.263,1.354) |
| Tuning $\lambda$ | 0.034 (0.022,0.043) | 0.064 (0.054,0.076) | 0.144 (0.134,0.149) | 0.324 (0.293,0.367) |
| $MSE_{ext}$ | 3.144 (3.095,3.206) | 3.176 (3.118,3.247) | 3.369 (3.256,3.491) | 4.244 (3.958,4.499) |
| $Optimism_{ext}$ | -0.159 (-0.433,0.149) | -0.152 (-0.424,0.148) | -0.180 (-0.480,0.123) | -0.323 (-0.687,0.049) |
| $Optimism_{int}$ | -0.107 (-0.136,-0.077) | -0.092 (-0.120,-0.064) | -0.074 (-0.101,-0.045) | -0.077 (-0.112,-0.044) |
| $MSE_{corrected}$ | 3.092 (2.819,3.397) | 3.116 (2.836,3.418) | 3.264 (2.966,3.594) | 3.998 (3.616,4.357) |
| $\beta_L P^*$ | 1.020 (1.014,1.025) | 1.049 (1.040,1.056) | 1.126 (1.111,1.141) | 1.297 (1.255,1.335) |
| | **MCAR** | | | |
| $MSE_{apparent}$ | 3.058 (2.729,3.410) | 3.105 (2.772,3.457) | 3.293 (2.941,3.667) | 4.071 (3.654,4.527) |
| $\beta_{LP}$ | 1.033 (1.024,1.040) | 1.054 (1.044,1.065) | 1.115 (1.098,1.133) | 1.257 (1.220,1.300) |
| Tuning $\lambda$ | 0.041 (0.035,0.367) | 0.069 (0.035,0.367) | 0.146 (0.035,0.367) | 0.331 (0.035,0.367) |
| $MSE_{ext}$ | 3.535 (3.351,3.805) | 3.633 (3.406,3.923) | 4.031 (3.698,4.446) | 5.456 (4.863,6.134) |
| $Optimism_{ext}$ | -0.478 (-0.875,-0.092) | -0.528 (-0.921,-0.148) | -0.738 (-1.161,-0.293) | -1.385 (-2.029,-0.866) |
| $Optimism_{int}$ | -0.070 (-0.156,0.008) | -0.051 (-0.137,0.024) | -0.030 (-0.114,0.039) | -0.052 (-0.122,0.009) |
| $MSE_{corrected}$ | 3.213 (2.851,3.644) | 3.237 (2.878,3.666) | 3.396 (3.009,3.848) | 4.192 (3.732,4.729) |
| $\beta_{LP}*$ | 1.018 (1.007,1.030) | 1.048 (1.035,1.060) | 1.129 (1.112,1.149) | 1.311 (1.266,1.355) |
| | **MAR** | | | |
| $MSE_{apparent}$ | 2.898 (2.598,3.204) | 2.945 (2.633,3.253) | 3.124 (2.792,3.478) | 3.867 (3.441,4.311) |
| $\beta_{LP}$ | 1.030 (1.022,1.038) | 1.051 (1.041,1.063) | 1.110 (1.094,1.127) | 1.251 (1.214,1.292) |
| Tuning $\lambda$ | 0.037 (0.031,0.328) | 0.065 (0.031,0.328) | 0.139 (0.031,0.328) | 0.319 (0.031,0.328) |
| $MSE_{ext}$ | 3.516 (3.312,3.777) | 3.595 (3.358,3.873) | 3.937 (3.608,4.305) | 5.245 (4.686,5.885) |
| $Optimism_{ext}$ | -0.618 (-1.026,-0.270) | -0.651 (-1.060,-0.279) | -0.813 (-1.269,-0.421) | -1.379 (-1.937,-0.842) |
| $Optimism_{int}$ | -0.164 (-0.258,-0.076) | -0.136 (-0.232,-0.049) | -0.110 (-0.190,-0.032) | -0.132 (-0.196,-0.061) |
| $MSE_{corrected}$ | 3.062 (2.714,3.390) | 3.081 (2.734,3.416) | 3.234 (2.873,3.596) | 3.999 (3.541,4.471) |
| $\beta_{LP^*}$ | 1.022 (1.008,1.039) | 1.044 (1.029,1.062) | 1.105 (1.087,1.126) | 1.249 (1.212,1.288) |

Table A.13: **Accuracy** simulation study results for **MissForest-Lasso** analysis with Harrell bootstrap validation: scenarios **S3** (assumption of moderation, without missing data) and **S4** (assumption of moderation, with missing data, complete outcome) based on 300 data sets of **20 variables** each (**n=1000**). Means of all estimates along with their corresponding 2.5th and 97.5th percentile values within parenthesis. The theoretical MSE is $1.74^2 = 3.028$

| **Estimates** | **Complete cases** | | | |
|---|---|---|---|---|
| | Best model | 1 SE tolerance | 3% tolerance | 15% tolerance |
| $MSE_{apparent}$ | 2.966 (2.697,3.251) | 3.009 (2.733,3.302) | 3.181 (2.885,3.500) | 3.911 (3.528,4.338) |
| $\beta_{LP}$ | 1.033 (1.026,1.040) | 1.052 (1.044,1.061) | 1.109 (1.097,1.125) | 1.240 (1.211,1.276) |
| Tuning $\lambda$ | 0.041 (0.031,0.049) | 0.066 (0.054,0.076) | 0.137 (0.119,0.149) | 0.298 (0.262,0.328) |
| $MSE_{ext}$ | 3.182 (3.114,3.268) | 3.217 (3.140,3.311) | 3.419 (3.292,3.555) | 4.312 (4.017,4.612) |
| $Optimism_{ext}$ | -0.216 (-0.487,0.092) | -0.208 (-0.475,0.101) | -0.238 (-0.541,0.063) | -0.401 (-0.765,-0.050) |
| $Optimism_{int}$ | -0.169 (-0.205,-0.138) | -0.143 (-0.177,-0.110) | -0.114 (-0.145,-0.084) | -0.114 (-0.148,-0.077) |
| $MSE_{corrected}$ | 3.135 (2.850,3.440) | 3.152 (2.863,3.457) | 3.295 (2.995,3.617) | 4.025 (3.638,4.472) |
| $\beta_{LP^*}$ | 1.019 (1.012,1.025) | 1.038 (1.029,1.046) | 1.097 (1.085,1.108) | 1.228 (1.200,1.259) |
| | **MCAR** | | | |
| $MSE_{apparent}$ | 3.406 (3.042,3.808) | 3.457 (3.082,3.866) | 3.661 (3.273,4.094) | 4.518 (4.047,5.057) |
| $\beta_{LP}$ | 1.033 (1.024,1.041) | 1.055 (1.044,1.065) | 1.116 (1.102,1.133) | 1.261 (1.225,1.303) |
| Tuning $\lambda$ | 0.043 (0.035,0.367) | 0.072 (0.035,0.367) | 0.154 (0.035,0.367) | 0.343 (0.035,0.367) |
| $MSE_{ext}$ | 3.291 (3.178,3.429) | 3.342 (3.220,3.509) | 3.627 (3.418,3.886) | 4.857 (4.357,5.419) |
| $Optimism_{ext}$ | 0.115 (-0.249,0.478) | 0.115 (-0.253,0.468) | 0.034 (-0.360,0.412) | -0.339 (-0.823,0.139) |
| $Optimism_{int}$ | -0.158 (-0.248,-0.072) | -0.123 (-0.210,-0.040) | -0.086 (-0.165,-0.004) | -0.105 (-0.186,-0.035) |
| $MSE_{corrected}$ | 3.562 (3.179,3.961) | 3.576 (3.190,3.981) | 3.743 (3.341,4.182) | 4.618 (4.134,5.157) |
| $\beta_{LP*}$ | 1.016 (1.007,1.026) | 1.038 (1.027,1.050) | 1.103 (1.087,1.119) | 1.249 (1.216,1.286) |
| | **MAR** | | | |
| $MSE_{apparent}$ | 3.306 (2.959,3.659) | 3.356 (2.998,3.726) | 3.554 (3.164,3.939) | 4.386 (3.905,4.909) |
| $\beta_{LP}$ | 1.032 (1.024,1.039) | 1.053 (1.044,1.062) | 1.114 (1.099,1.130) | 1.258 (1.222,1.300) |
| Tuning $\lambda$ | 0.042 (0.035,0.367) | 0.070 (0.035,0.367) | 0.150 (0.035,0.367) | 0.338 (0.035,0.367) |
| $MSE_{ext}$ | 3.295 (3.183,3.467) | 3.339 (3.201,3.523) | 3.602 (3.389,3.866) | 4.764 (4.276,5.299) |
| $Optimism_{ext}$ | 0.011 (-0.409,0.405) | 0.016 (-0.383,0.407) | -0.047 (-0.495,0.357) | -0.377 (-0.852,0.135) |
| $Optimism_{int}$ | -0.159 (-0.258,-0.056) | -0.126 (-0.218,-0.025) | -0.091 (-0.180,0.003) | -0.108 (-0.190,-0.026) |
| $MSE_{corrected}$ | 3.610 (2.886,4.346) | 3.670 (2.929,4.429) | 3.847 (3.088,4.641) | 4.837 (3.938,5.964) |
| $\beta_{LP^*}$ | 1.018 (1.007,1.028) | 1.039 (1.028,1.050) | 1.103 (1.088,1.119) | 1.249 (1.213,1.281) |

Table A.14: **Accuracy** simulation study results for **MissForest-Lasso** analysis with Harrell bootstrap validation: scenarios **S3** (assumption of moderation, without missing data) and **S5** (assumption of moderation, with missing data also in the outcome) based on 300 data sets of **20 variables** each (**n=1000**). Means of all estimates along with their corresponding 2.5th and 97.5th percentile values within parenthesis. The theoretical MSE is $1.74^2 = 3.028$

| **Estimates** | **Complete cases** | | | |
|---|---|---|---|---|
| | Best model | 1 SE tolerance | 3% tolerance | 15% tolerance |
| $MSE_{apparent}$ | 2.764 (2.232,3.313) | 2.873 (2.328,3.483) | 3.076 (2.485,3.729) | 3.939 (3.121,4.889) |
| $\beta_{LP}$ | 1.062 (1.044,1.079) | 1.089 (1.069,1.112) | 1.132 (1.102,1.165) | 1.259 (1.204,1.332) |
| Tuning $\lambda$ | 0.077 (0.054,0.095) | 0.111 (0.085,0.134) | 0.165 (0.134,0.209) | 0.328 (0.262,0.410) |
| $MSE_{ext}$ | 3.482 (3.266,3.794) | 3.557 (3.313,3.914) | 3.759 (3.436,4.214) | 4.799 (4.070,5.835) |
| $Optimism_{ext}$ | -0.719 (-1.300,-0.093) | -0.684 (-1.283,-0.050) | -0.683 (-1.281,0.017) | -0.860 (-1.706,-0.064) |
| $Optimism_{int}$ | -0.646 (-0.806,-0.521) | -0.578 (-0.731,-0.462) | -0.523 (-0.665,-0.420) | -0.486 (-0.620,-0.373) |
| $MSE_{corrected}$ | 3.410 (2.788,4.131) | 3.451 (2.826,4.194) | 3.598 (2.937,4.384) | 4.425 (3.541,5.495) |
| $\beta_{LP*}$ | 1.025 (1.012,1.036) | 1.052 (1.035,1.069) | 1.092 (1.067,1.118) | 1.216 (1.167,1.273) |
| | **MCAR** | | | |
| $MSE_{apparent}$ | 3.132 (2.448,3.872) | 3.283 (2.553,4.032) | 3.551 (2.764,4.388) | 4.761 (3.686,5.978) |
| $\beta_{LP}$ | 1.075 (1.050,1.102) | 1.108 (1.077,1.146) | 1.161 (1.118,1.216) | 1.343 (1.255,1.470) |
| Tuning $\lambda$ | 0.090 (0.068,0.513) | 0.134 (0.068,0.513) | 0.204 (0.068,0.513) | 0.455 (0.068,0.513) |
| $MSE_{ext}$ | 4.365 (3.743,5.132) | 4.619 (3.895,5.522) | 5.106 (4.163,6.226) | 7.083 (5.582,8.833) |
| $Optimism_{ext}$ | -1.234 (-2.021,-0.334) | -1.336 (-2.189,-0.380) | -1.555 (-2.575,-0.506) | -2.323 (-3.692,-0.931) |
| $Optimism_{int}$ | -0.831 (-1.122,-0.582) | -0.737 (-0.979,-0.522) | -0.659 (-0.882,-0.461) | -0.547 (-0.760,-0.372) |
| $MSE_{corrected}$ | 4.174 (3.360,5.101) | 4.218 (3.377,5.188) | 4.407 (3.527,5.440) | 5.519 (4.345,6.897) |
| $\beta_{LP*}$ | 1.028 (1.004,1.050) | 1.066 (1.039,1.097) | 1.118 (1.084,1.160) | 1.297 (1.220,1.399) |
| | **MAR** | | | |
| $MSE_{apparent}$ | 2.910 (2.233,3.653) | 3.046 (2.351,3.860) | 3.288 (2.519,4.152) | 4.402 (3.417,5.579) |
| $\beta_{LP}$ | 1.068 (1.046,1.094) | 1.100 (1.074,1.135) | 1.149 (1.113,1.194) | 1.321 (1.241,1.450) |
| Tuning $\lambda$ | 0.081 (0.061,0.513) | 0.122 (0.061,0.513) | 0.185 (0.061,0.513) | 0.415 (0.061,0.513) |
| $MSE_{ext}$ | 4.260 (3.637,5.099) | 4.461 (3.738,5.452) | 4.872 (3.961,6.137) | 6.680 (5.168,8.522) |
| $Optimism_{ext}$ | -1.351 (-2.478,-0.554) | -1.415 (-2.543,-0.582) | -1.584 (-2.816,-0.624) | -2.278 (-3.766,-0.978) |
| $Optimism_{int}$ | -0.795 (-1.073,-0.570) | -0.711 (-0.960,-0.507) | -0.645 (-0.860,-0.447) | -0.559 (-0.744,-0.374) |
| $MSE_{corrected}$ | 4.087 (3.292,5.094) | 4.132 (3.344,5.124) | 4.315 (3.466,5.338) | 5.395 (4.309,6.784) |
| $\beta_{LP^*}$ | 1.027 (1.002,1.053) | 1.063 (1.034,1.095) | 1.112 (1.075,1.155) | 1.281 (1.212,1.379) |

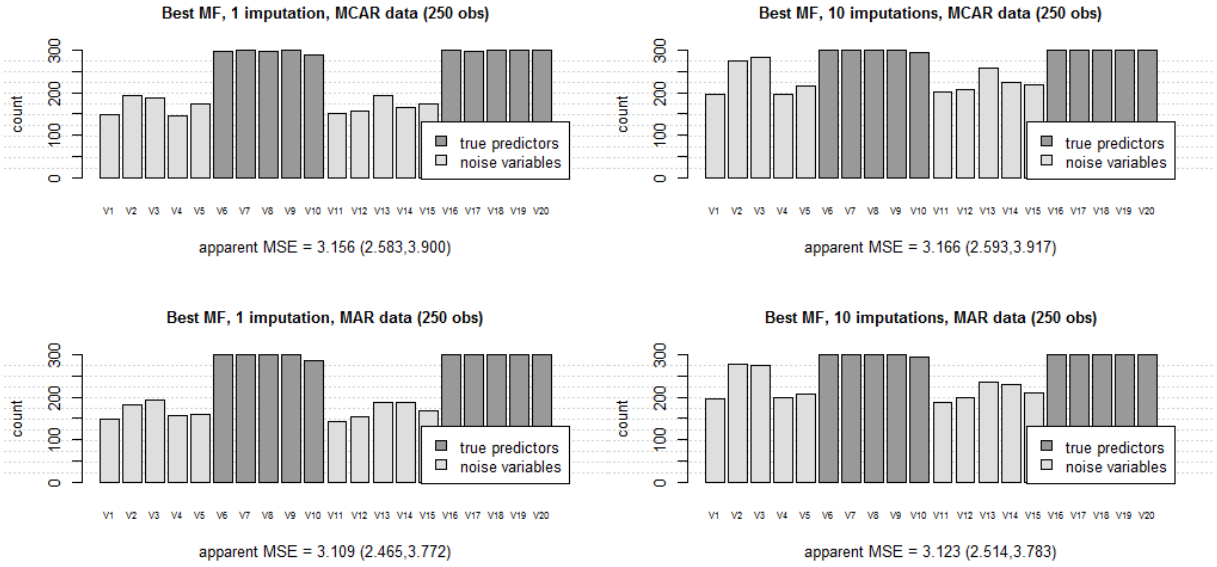
Table A.15: **Accuracy** simulation study results for **MissForest-Lasso** analysis with Harrell (1996) bootstrap validation: scenarios **S3** (assumption of moderation, complete data) and **S5** (assumption of moderation, missing data also in the outcome), based on 300 data sets of **100 variables** each (**n=500**) with between-covariate **correlation of 0.8**. Means of all estimates along with their corresponding 2.5th and 97.5th percentile values within parenthesis. The theoretical MSE is 1.

| **Estimates** | **Complete cases** | | | |
|---|---|---|---|---|
| | Best model | 1 SE tolerance | 3% tolerance | 15% tolerance |
| $\text{MSE}_{apparent}$ | 1.000 (0.852,1.146) | 1.053 (0.916,1.215) | 1.322 (1.153,1.501) | 1.528 (1.337,1.738) |
| $\beta_{LP}$ | 1.054 (1.044,1.065) | 1.063 (1.050,1.077) | 1.127 (1.099,1.155) | 1.188 (1.145,1.242) |
| Tuning $\lambda$ | 0.044 (0.031,0.054) | 0.141 (0.107,0.187) | 0.056 (0.039,0.068) | 0.224 (0.167,0.293) |
| $\text{MSE}_{ext}$ | 1.112 (1.041,1.183) | 1.128 (1.054,1.212) | 1.320 (1.188,1.465) | 1.516 (1.336,1.718) |
| $\text{Optimism}_{ext}$ | -0.112 (-0.255,0.044) | -0.075 (-0.206,0.075) | 0.002 (-0.159,0.160) | 0.012 (-0.179,0.203) |
| $\text{Optimism}_{int}$ | -0.302 (-0.351,-0.259) | -0.255 (-0.303,-0.211) | -0.122 (-0.160,-0.090) | -0.090 (-0.120,-0.063) |
| $\text{MSE}_{corrected}$ | 1.302 (1.134,1.491) | 1.308 (1.152,1.494) | 1.444 (1.263,1.629) | 1.617 (1.414,1.832) |
| $\beta_{LP^*}$ | 1.016 (1.010,1.024) | 1.026 (1.019,1.036) | 1.078 (1.061,1.099) | 1.122 (1.096,1.156) |
| **MCAR** | | | | |
| $\text{MSE}_{apparent}$ | 1.122 (0.948,1.308) | 1.179 (0.995,1.367) | 1.476 (1.263,1.716) | 1.699 (1.459,1.957) |
| $\beta_{LP}$ | 1.053 (1.043,1.067) | 1.064 (1.051,1.082) | 1.147 (1.110,1.195) | 1.239 (1.167,1.345) |
| Tuning $\lambda$ | 0.045 (0.035,0.367) | 0.061 (0.035,0.367) | 0.182 (0.035,0.367) | 0.305 (0.035,0.367) |
| $\text{MSE}_{ext}$ | 1.310 (1.178,1.469) | 1.337 (1.193,1.526) | 1.657 (1.408,1.972) | 1.964 (1.623,2.367) |
| $\text{Optimism}_{ext}$ | -0.189 (-0.415,0.057) | -0.157 (-0.385,0.079) | -0.182 (-0.474,0.095) | -0.265 (-0.602,0.074) |
| $\text{Optimism}_{int}$ | -0.368 (-0.438,-0.306) | -0.307 (-0.378,-0.246) | -0.143 (-0.189,-0.098) | -0.103 (-0.144,-0.065) |
| $\text{MSE}_{corrected}$ | 1.490 (1.290,1.729) | 1.486 (1.280,1.720) | 1.619 (1.405,1.872) | 1.802 (1.567,2.059) |
| $\beta_{LP^*}$ | 1.004 (0.993,1.015) | 1.016 (1.005,1.028) | 1.078 (1.059,1.103) | 1.142 (1.105,1.194) |
| **MAR** | | | | |
| $\text{MSE}_{apparent}$ | 1.145 (0.977,1.322) | 1.202 (1.021,1.387) | 1.504 (1.281,1.734) | 1.742 (1.472,2.012) |
| $\beta_{LP}$ | 1.054 (1.044,1.065) | 1.065 (1.053,1.080) | 1.143 (1.108,1.197) | 1.236 (1.165,1.360) |
| Tuning $\lambda$ | 0.045 (0.035,0.367) | 0.060 (0.035,0.367) | 0.174 (0.035,0.367) | 0.300 (0.035,0.367) |
| $\text{MSE}_{ext}$ | 1.305 (1.183,1.455) | 1.327 (1.195,1.499) | 1.635 (1.413,1.938) | 1.950 (1.633,2.325) |
| $\text{Optimism}_{ext}$ | -0.160 (-0.384,0.051) | -0.125 (-0.347,0.086) | -0.131 (-0.417,0.131) | -0.208 (-0.559,0.092) |
| $\text{Optimism}_{int}$ | -0.378 (-0.448,-0.311) | -0.316 (-0.381,-0.253) | -0.148 (-0.193,-0.106) | -0.109 (-0.153,-0.070) |
| $\text{MSE}_{corrected}$ | 1.620 (1.350,1.911) | 1.628 (1.357,1.923) | 1.724 (1.456,2.040) | 1.865 (1.575,2.191) |
| $\beta_{LP^*}$ | 1.005 (0.992,1.019) | 1.018 (1.003,1.032) | 1.081 (1.058,1.109) | 1.142 (1.105,1.197) |

Table A.16: **Accuracy** simulation study results for **MissForest-Elasticnet** analysis with Harrell (1996) bootstrap validation: scenarios **S3** (assumption of moderation, complete data) and **S5** (assumption of moderation, missing data also in the outcome), based on 300 data sets of **100 variables** each (**n=500**) with between-covariate **correlation of 0.8**. Means of all estimates along with their corresponding 2.5th and 97.5th percentile values within parenthesis. The theoretical MSE is $1$.

| **Estimates** | **Complete cases** | | | |
|---|---|---|---|---|
| | Best model | 1 SE tolerance | 3% tolerance | 15% tolerance |
| MSE$_{apparent}$ | 0.998 (0.842,1.148) | 1.092 (0.935,1.262) | 2.134 (1.806,2.549) | 2.944 (2.384,3.617) |
| $\beta_{LP}$ | 1.054 (1.044,1.066) | 1.073 (1.055,1.095) | 1.435 (1.274,1.677) | 2.140 (1.595,3.105) |
| Tuning $\alpha$ | 0.897 (0.800,0.900) | 0.900 (0.900,0.900) | 0.900 (0.900,0.900) | 0.900 (0.900,0.900) |
| Tuning $\lambda$ | 0.048 (0.032,0.065) | 0.546 (0.357,0.739) | 0.075 (0.051,0.106) | 0.976 (0.739,1.199) |
| MSE$_{ext}$ | 1.113 (1.040,1.184) | 1.150 (1.062,1.261) | 2.151 (1.733,2.541) | 3.018 (2.369,3.888) |
| Optimism$_{ext}$ | -0.115 (-0.258,0.049) | -0.058 (-0.193,0.102) | -0.017 (-0.285,0.264) | -0.074 (-0.468,0.318) |
| Optimism$_{int}$ | -0.313 (-0.373,-0.261) | -0.209 (-0.264,-0.159) | -0.092 (-0.137,-0.060) | -0.077 (-0.114,-0.046) |
| MSE$_{corrected}$ | 1.311 (1.134,1.491) | 1.302 (1.152,1.494) | 2.226 (1.263,1.629) | 3.021 (1.414,1.832) |
| $\beta_{LP^*}$ | 1.018 (1.011,1.026) | 1.047 (1.033,1.066) | 1.202 (1.130,1.304) | 1.379 (1.223,1.647) |
| | | **MCAR** | | |
| MSE$_{apparent}$ | 1.118 (0.951,1.302) | 1.324 (1.126,1.570) | 2.526 (1.789,3.370) | 4.270 (2.267,5.634) |
| $\beta_{LP}$ | 1.054 (1.043,1.067) | 1.102 (1.069,1.166) | 2.006 (1.302,3.131) | 25.617 (1.718,164.182) |
| Tuning $\lambda$ | 0.056 (0.040,0.106) | 0.142 (0.083,0.280) | 0.956 (0.655,1.199) | 1.834 (1.199,2.482) |
| MSE$_{ext}$ | 1.316 (1.177,1.483) | 1.488 (1.262,1.869) | 2.963 (2.222,4.026) | 4.841 (2.904,5.854) |
| Optimism$_{ext}$ | -0.198 (-0.433,0.037) | -0.164 (-0.451,0.083) | -0.437 (-0.957,0.035) | -0.571 (-1.280,0.077) |
| Optimism$_{int}$ | -0.407 (-0.482,-0.339) | -0.314 (-0.394,-0.228) | -0.159 (-0.219,-0.106) | -0.115 (-0.162,-0.069) |
| MSE$_{corrected}$ | 1.525 (1.329,1.771) | 1.638 (1.375,1.922) | 2.685 (1.961,3.541) | 4.385 (2.393,5.757) |
| $\beta_{LP^*}$ | 1.005 (0.994,1.016) | 1.029 (1.013,1.049) | 1.128 (1.069,1.232) | 1.862 (1.121,2.481) |
| | | **MAR** | | |
| MSE$_{apparent}$ | 1.145 (0.967,1.338) | 1.327 (1.102,1.598) | 2.566 (1.970,3.322) | 4.376 (2.931,5.519) |
| $\beta_{LP}$ | 1.055 (1.044,1.067) | 1.096 (1.067,1.148) | 2.011 (1.445,3.233) | NA (2.231,117.110) |
| Tuning $\alpha$ | 0.051 (0.040,0.065) | 0.117 (0.065,0.220) | 0.908 (0.580,1.199) | 1.779 (1.199,2.482) |
| Tuning $\lambda$ | 0.041 (0.032,0.063) | 0.291 (0.176,0.449) | 0.067 (0.044,0.102) | 0.586 (0.348,0.928) |
| MSE$_{ext}$ | 1.307 (1.187,1.476) | 1.442 (1.240,1.815) | 2.950 (2.254,4.022) | 4.913 (3.303,5.836) |
| Optimism$_{ext}$ | -0.163 (-0.429,0.065) | -0.115 (-0.374,0.126) | -0.384 (-0.889,0.050) | -0.537 (-1.277,0.146) |
| Optimism$_{int}$ | -0.419 (-0.492,-0.344) | -0.317 (-0.404,-0.231) | -0.161 (-0.217,-0.111) | -0.119 (-0.161,-0.083) |
| MSE$_{corrected}$ | 1.564 (1.324,1.801) | 1.644 (1.342,1.979) | 2.727 (2.126,3.512) | 4.495 (3.047,5.655) |
| $\beta_{LP^*}$ | -0.419 (-0.492,-0.344) | -0.317 (-0.404,-0.231) | -0.161 (-0.217,-0.111) | -0.119 (-0.161,-0.083) |

Figure A.4: Comparison of inclusion frequency of the variables in 300 simulated 20-covariate datasets (**250 obs**) for the best **MissForest-LASSO** models with bootstrap tuning with single imputation VS ten imputations. Here we can see how MissForest outperformed MICE by not selecting all variables almost always.




## A.3 Simulation result figures: method comparison

Figure A.5: **Optimism-corrected MSE** estimates from 4 methods run on 300 simulated **20-covariate** datasets with **1000 observations** for scenarios **S1** (without missing data) and **S2** (with missing data). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME), MissForest-Lasso (MF) and MissForest-Random Forests (MR). ML, ME and MF estimated MSEs are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S1 (first plot from the left), the Lasso (L), Elasticnet (E) and Random Forest (RF) corrected MSEs are shown.

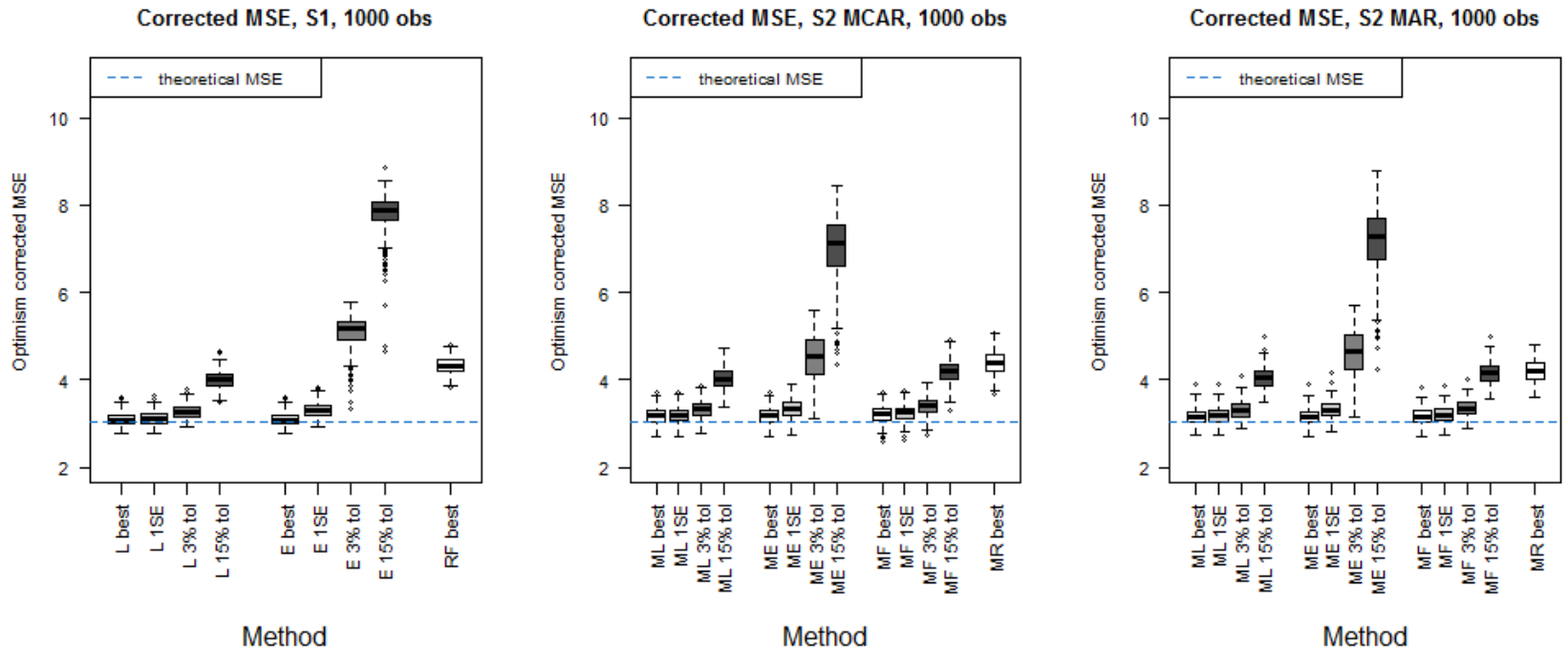


Figure A.6: **Optimism-corrected MSE** estimates from 4 methods run on 300 simulated **20-covariate** datasets with **1000 observations** for scenarios **S3** (assumption of moderation, without missing data) and **S4** (assumption of moderation, with missing data). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME), MissForest-Lasso (MF) and MissForest-Random Forests (MR). ML, ME and MF estimated MSEs are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S3 (first plot from the left), the Lasso (L), Elasticnet (E) and Random Forest (RF) corrected MSEs are shown.

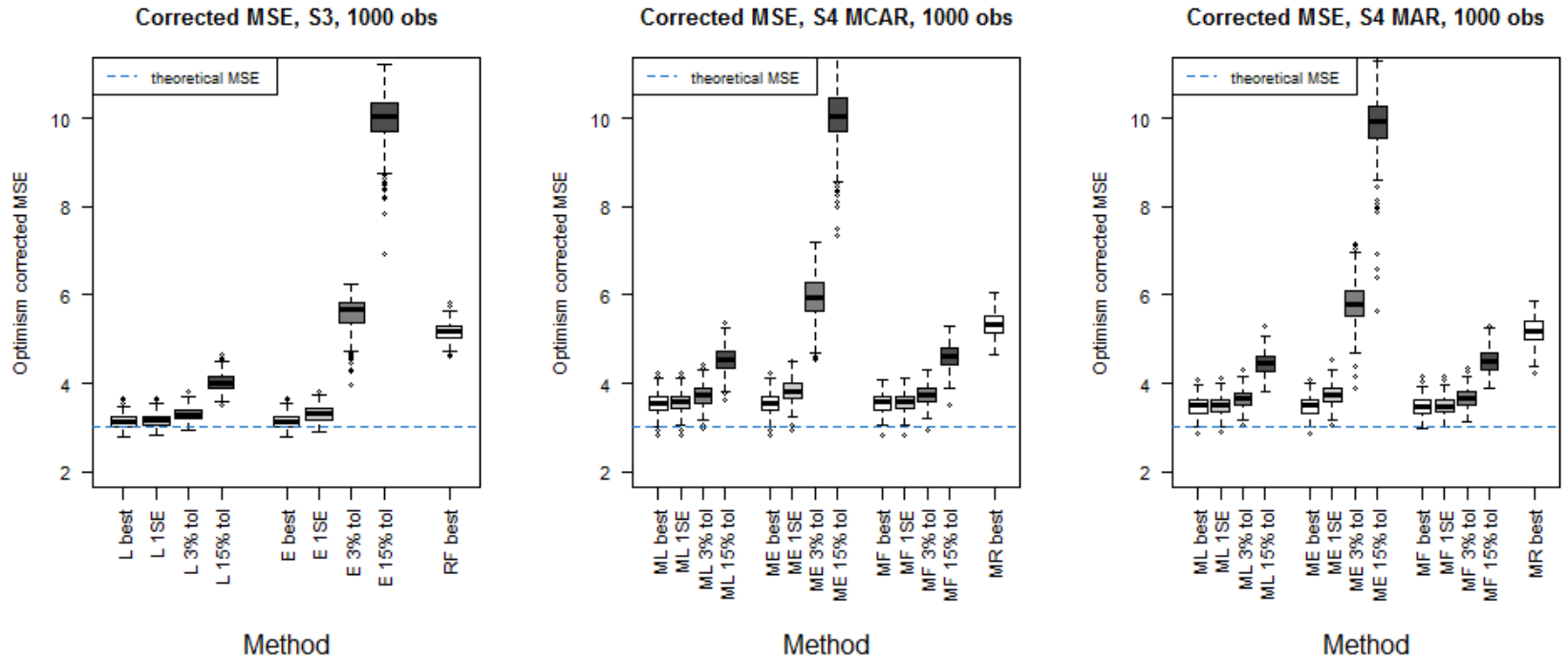


Figure A.7: **Optimism-corrected MSE** estimates from 4 methods run on 300 simulated **20-covariate** datasets with **1000 observations** for scenarios **S3** (assumption of moderation, without missing data) and **S5** (assumption of moderation, with missing data also in the outcome). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME), MissForest-Lasso (MF) and MissForest-Random Forests (MR). ML, ME and MF estimated MSEs are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S3 (first plot from the left), the Lasso (L), Elasticnet (E) and Random Forest (RF) corrected MSEs are shown.

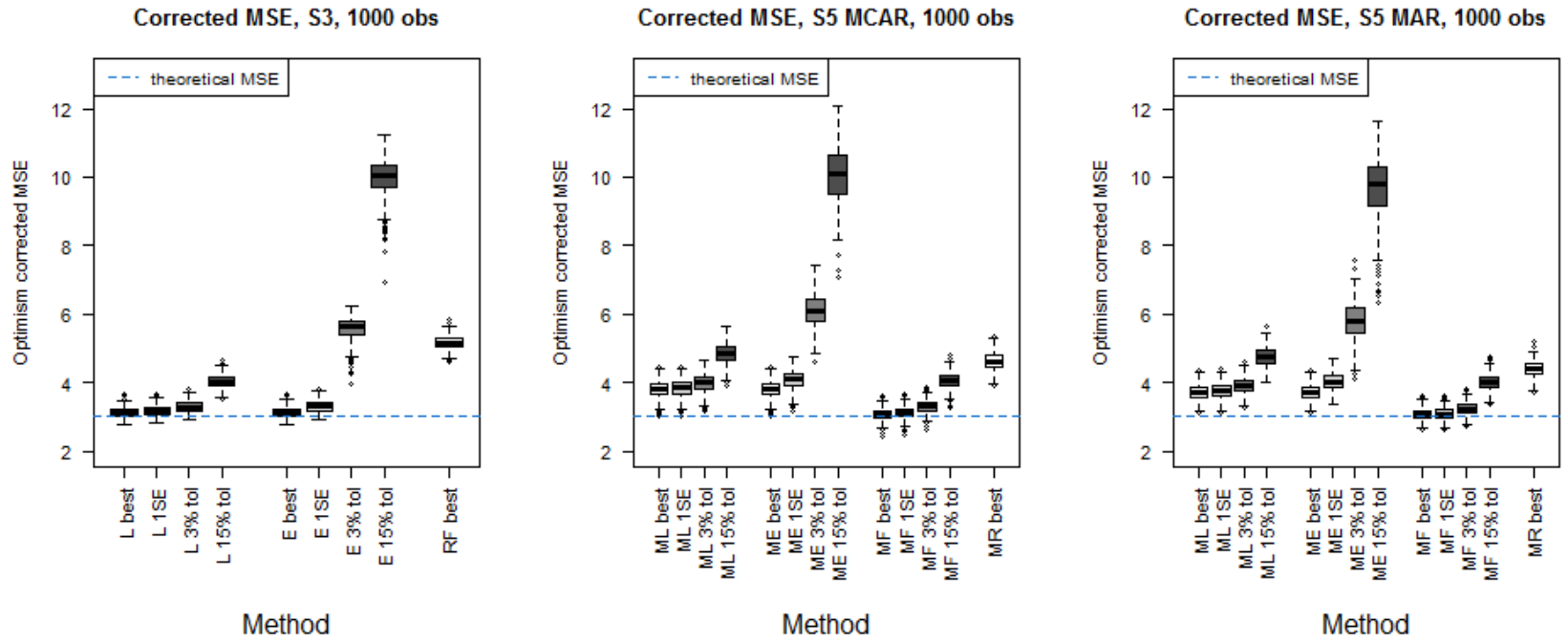


Figure A.8: **Calibration slope** $\beta_{LP}$ estimates for 4 methods run on 300 simulated **20-covariate** datasets with **1000 observations** for scenarios **S1** (without missing data) and **S2** (with missing data). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME), MissForest-Lasso (MF) and MissForest-Random Forests (MR). ML, ME and MF estimated calibration slopes are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S1 (first plot from the left), the Lasso (L), Elasticnet (E) and Random Forest (RF) calibration slopes are shown.

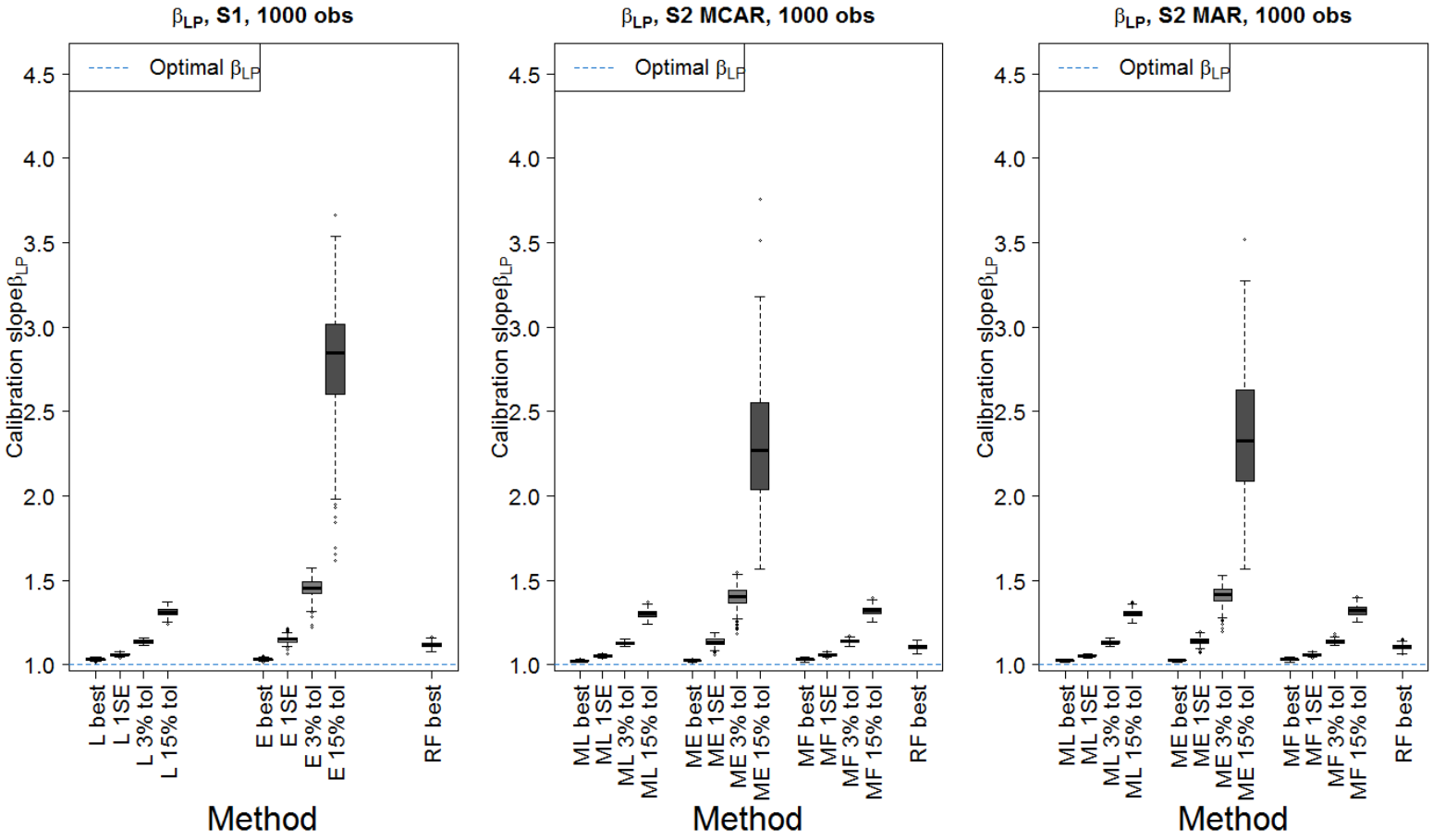


Figure A.9: **Calibration slope** $\beta_{LP}$ estimates for 4 methods run on 300 simulated **20-covariate** datasets with **1000 observations** for scenarios **S3** (assumption of moderation, without missing data) and **S4** (assumption of moderation, with missing data). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME), MissForest-Lasso (MF) and MissForest-Random Forests (MR). ML, ME and MF estimated calibration slopes are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S3 (first plot from the left), the Lasso (L), Elasticnet (E) and Random Forest (RF) calibration slopes are shown.

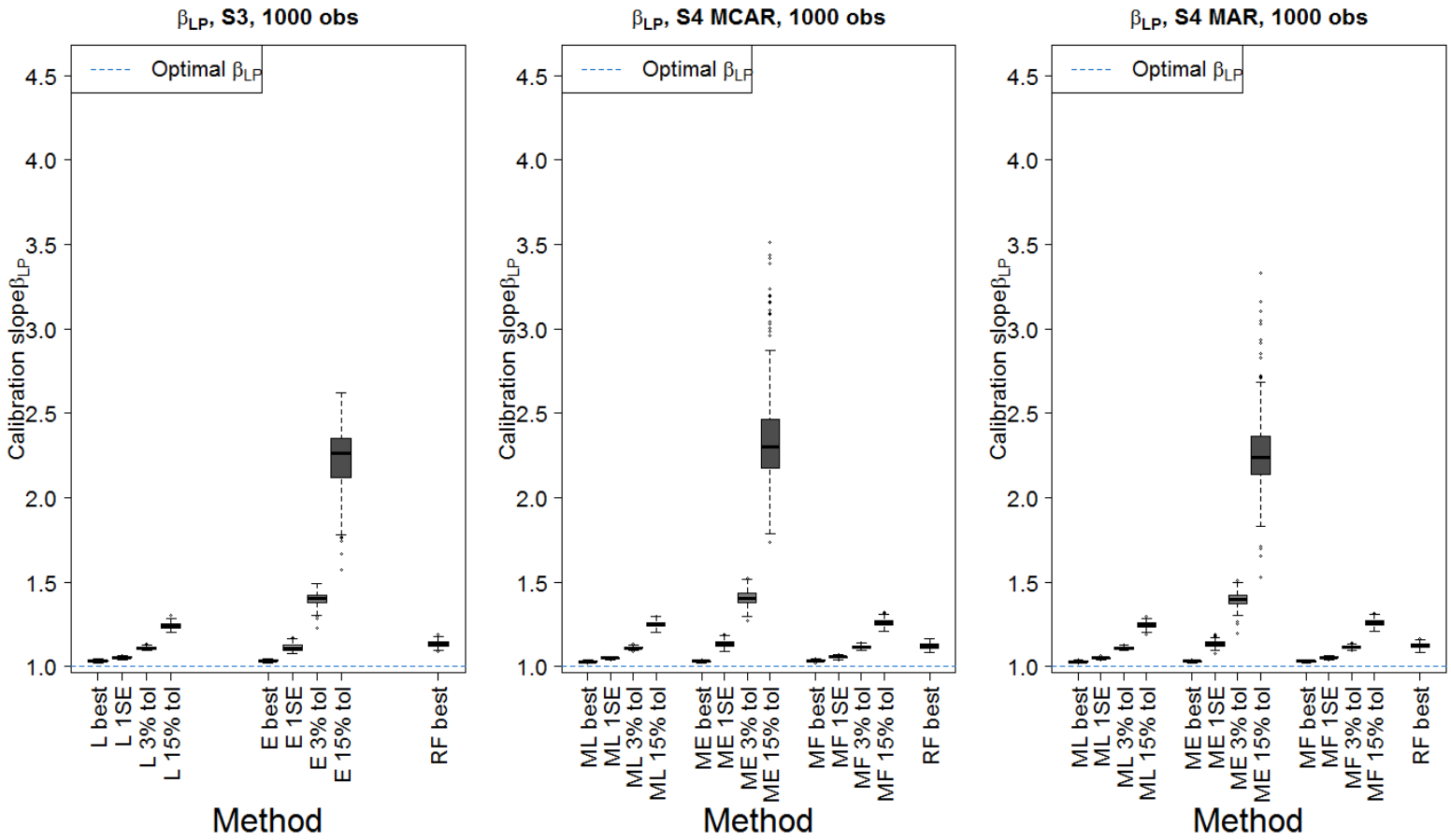


Figure A.10: **Calibration slope** $\beta_{LP}$ estimates for 4 methods run on 300 simulated **20-covariate** datasets with **1000 observations** for scenarios **S3** (assumption of moderation, without missing data) and **S5** (assumption of moderation, missing data also in the outcome). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME), MissForest-Lasso (MF) and MissForest-Random Forests (MR). ML, ME and MF estimated calibration slopes are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S3 (first plot from the left), the Lasso (L), Elasticnet (E) and Random Forest (RF) calibration slopes are shown.

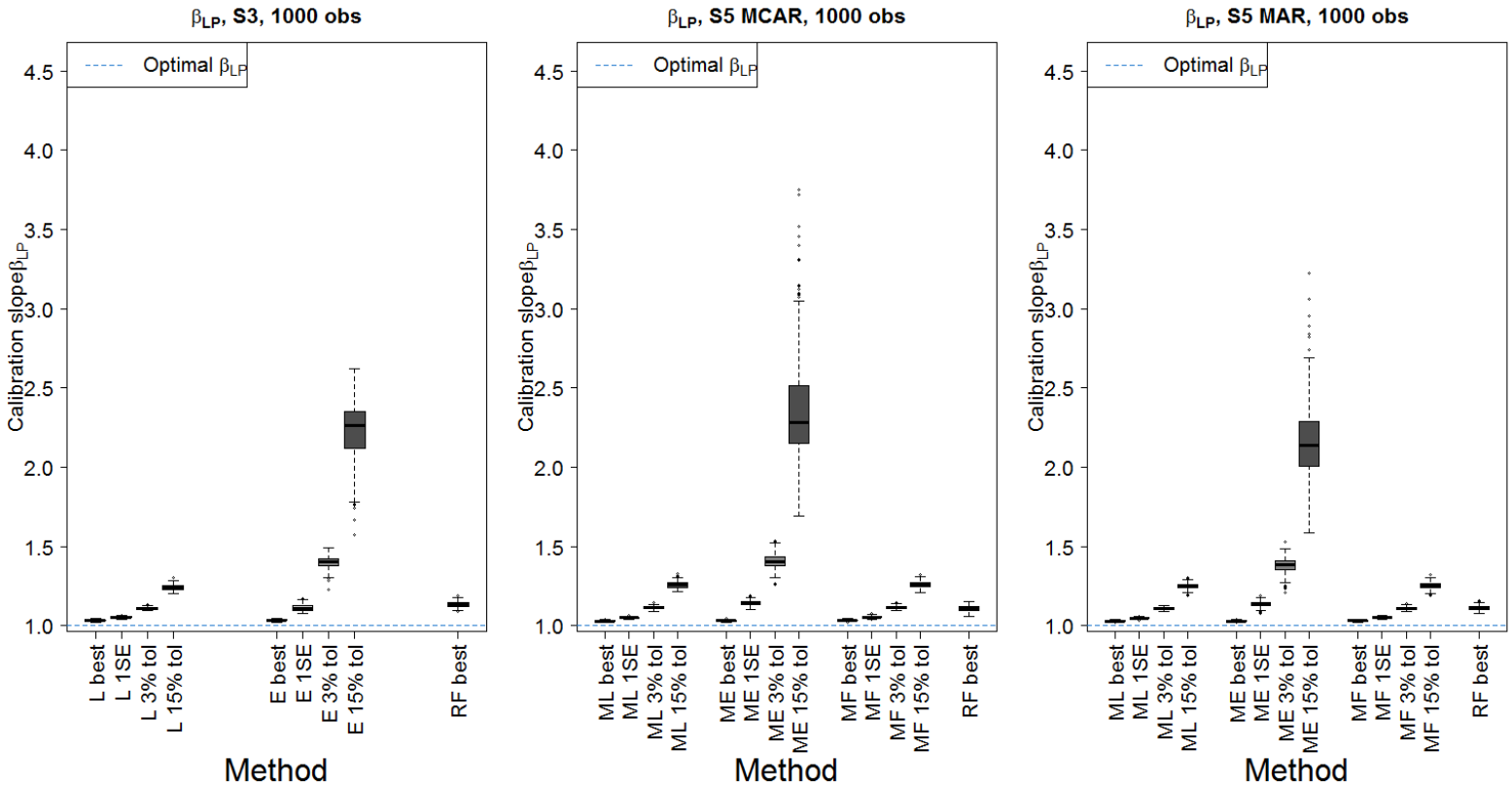


Figure A.11: Average **internal and external MSE optimism** estimates with 2.5th and 97.5th percentiles for 4 methods run on 300 simulated **20-covariate** datasets with **1000 observations** for scenarios **S1** (without missing data) and **S2** (with missing data). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME), MissForest-Lasso (MF) and MissForest-Random Forests (MR). ML, ME and MF estimated internal and external MSE optimism are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S1 (first plot from the left), the Lasso (L), Elasticnet (E) and Random Forest (RF) optimism estimates are shown.

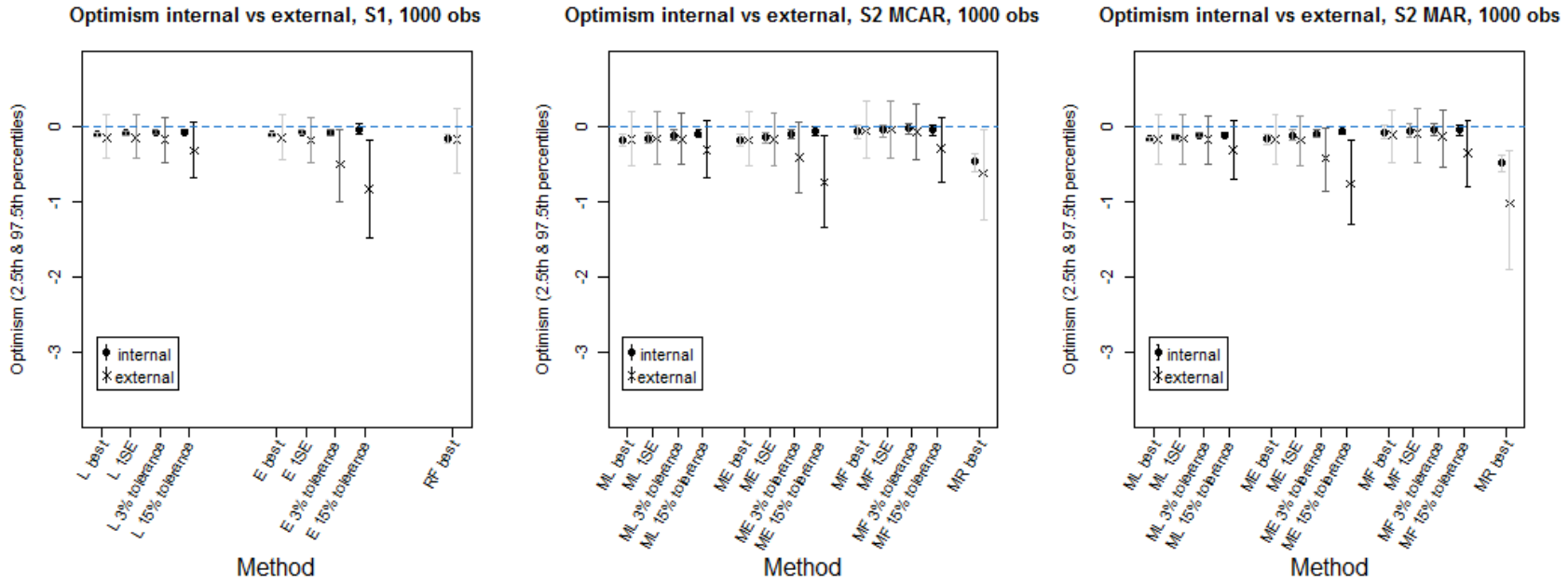


Figure A.12: Average **internal and external MSE optimism** estimates with 2.5th and 97.5th percentiles for 4 methods run on 300 simulated **20-covariate** datasets with **1000 observations** for scenarios **S3** (assumption of moderation, without missing data) and **S4** (assumption of moderation, with missing data). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME), MissForest-Lasso (MF) and MissForest-Random Forests (MR). ML, ME and MF estimated internal and external MSE optimism are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S3 (first plot from the left), the Lasso (L), Elasticnet (E) and Random Forest (RF) optimism estimates are shown.

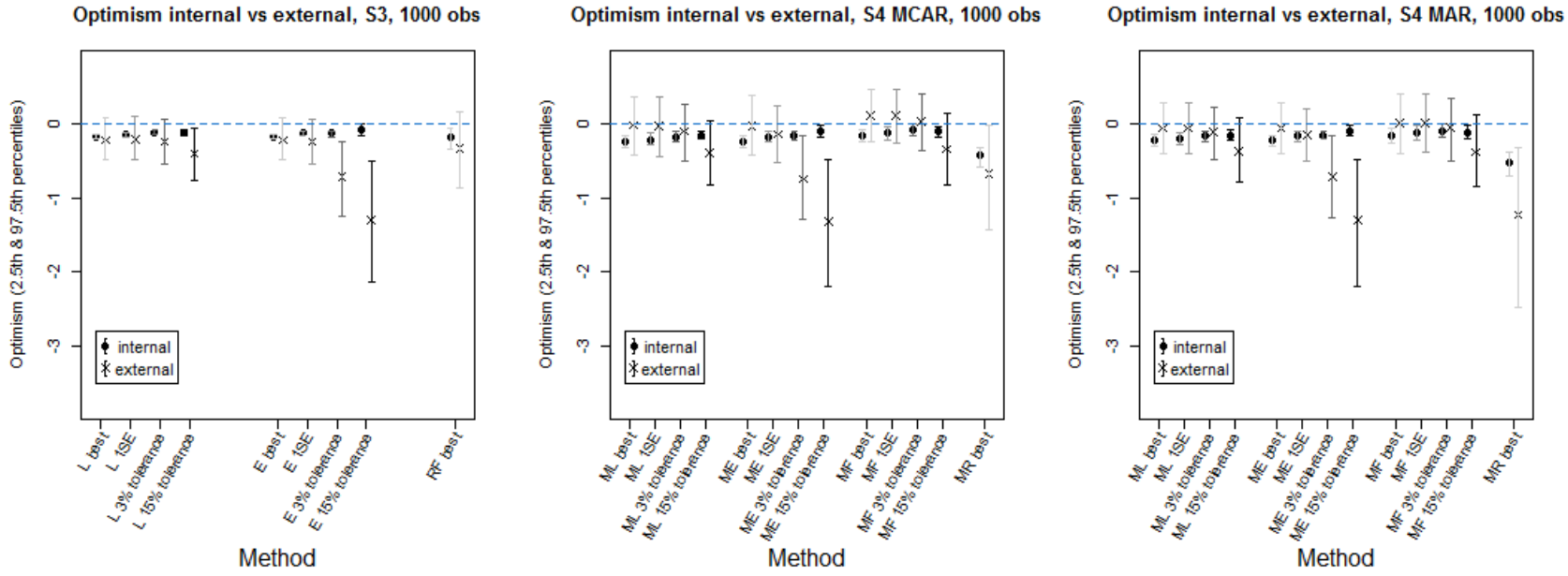


Figure A.13: Average **internal and external MSE optimism** estimates with 2.5th and 97.5th percentiles for 4 methods run on 300 simulated **20-covariate** datasets with **1000 observations** for scenarios **S3** (assumption of moderation, without missing data) and **S5** (assumption of moderation, with missing data also in the outcome). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME), MissForest-Lasso (MF) and MissForest-Random Forests (MR). ML, ME and MF estimated internal and external MSE optimism are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S3 (first plot from the left), the Lasso (L), Elasticnet (E) and Random Forest (RF) optimism estimates are shown.

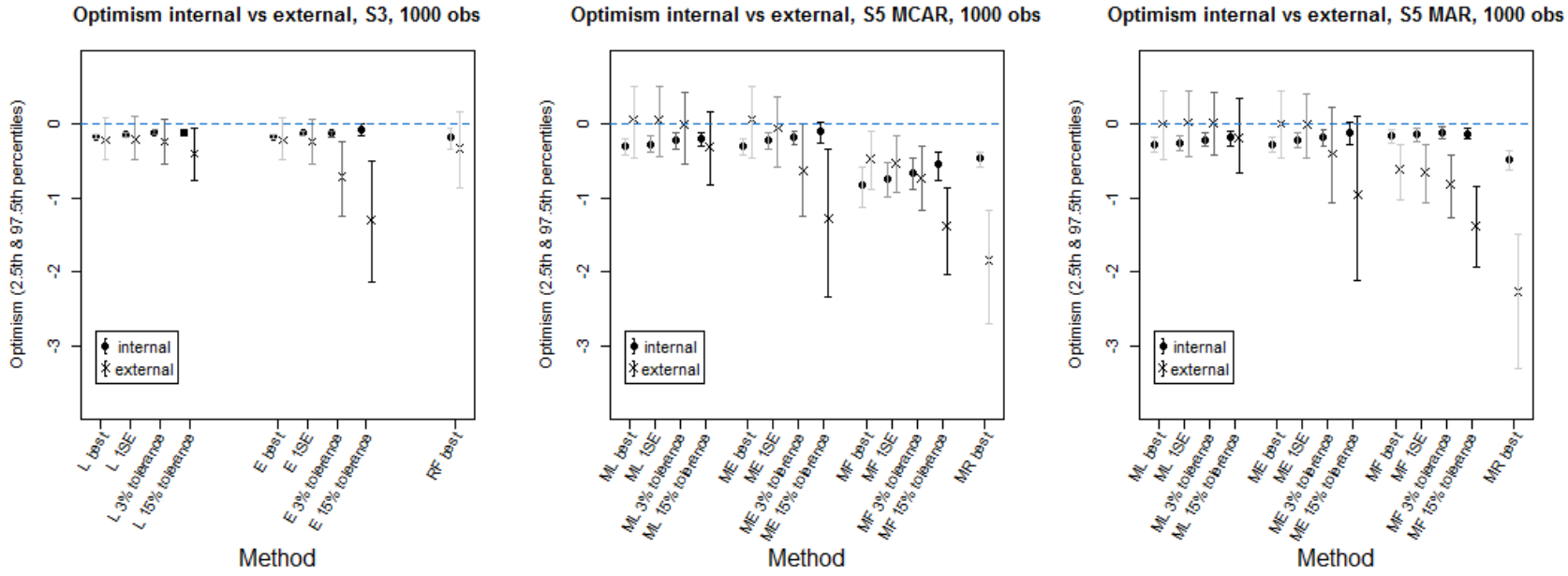


Figure A.14: Average percentage of **true predictors (TP) selected among the actual TP** (SEN) estimates with 2.5th and 97.5th percentiles from 3 methods run on 300 simulated **20-covariate** datasets with **1000 observations** for scenarios **S1** (without missing data) and **S2** (with missing data). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME) and MissForest-Lasso (MF). ML, ME and MF estimated percentages of TP selected among the actual TP variables are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S1 (first plot from the left), the Lasso (L) and the Elasticnet (E) estimates are shown.

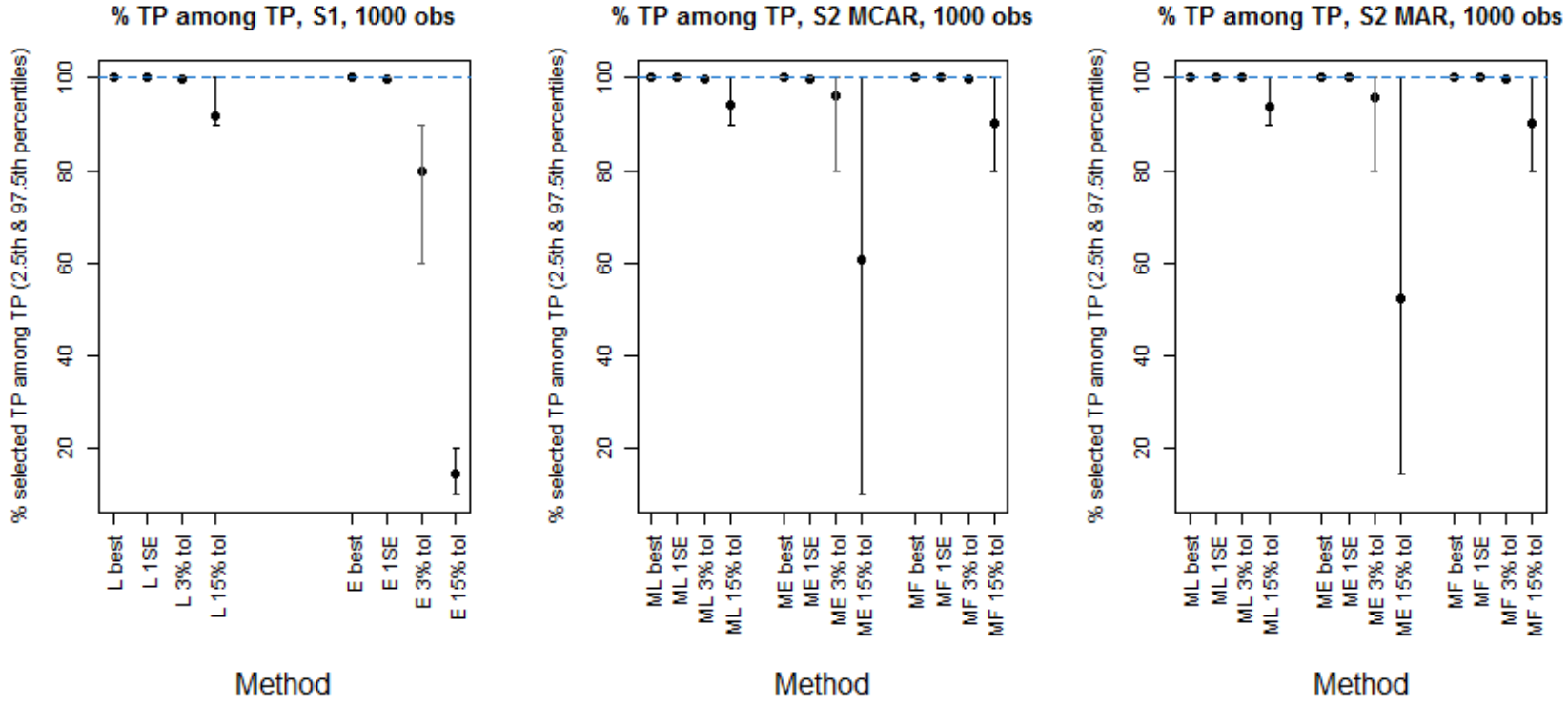


Figure A.15: Average percentage of **true predictors (TP) selected among the actual TP** (SEN) estimates with 2.5th and 97.5th percentiles from 3 methods run on 300 simulated **20-covariate** datasets with **1000 observations** for scenarios **S3** (assumption of moderation, without missing data) and **S4** (assumption of moderation, with missing data). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME) and MissForest-Lasso (MF). ML, ME and MF estimated percentages of TP selected among the actual TP variables are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S3 (first plot from the left), the Lasso (L) and the Elasticnet (E) estimates are shown.

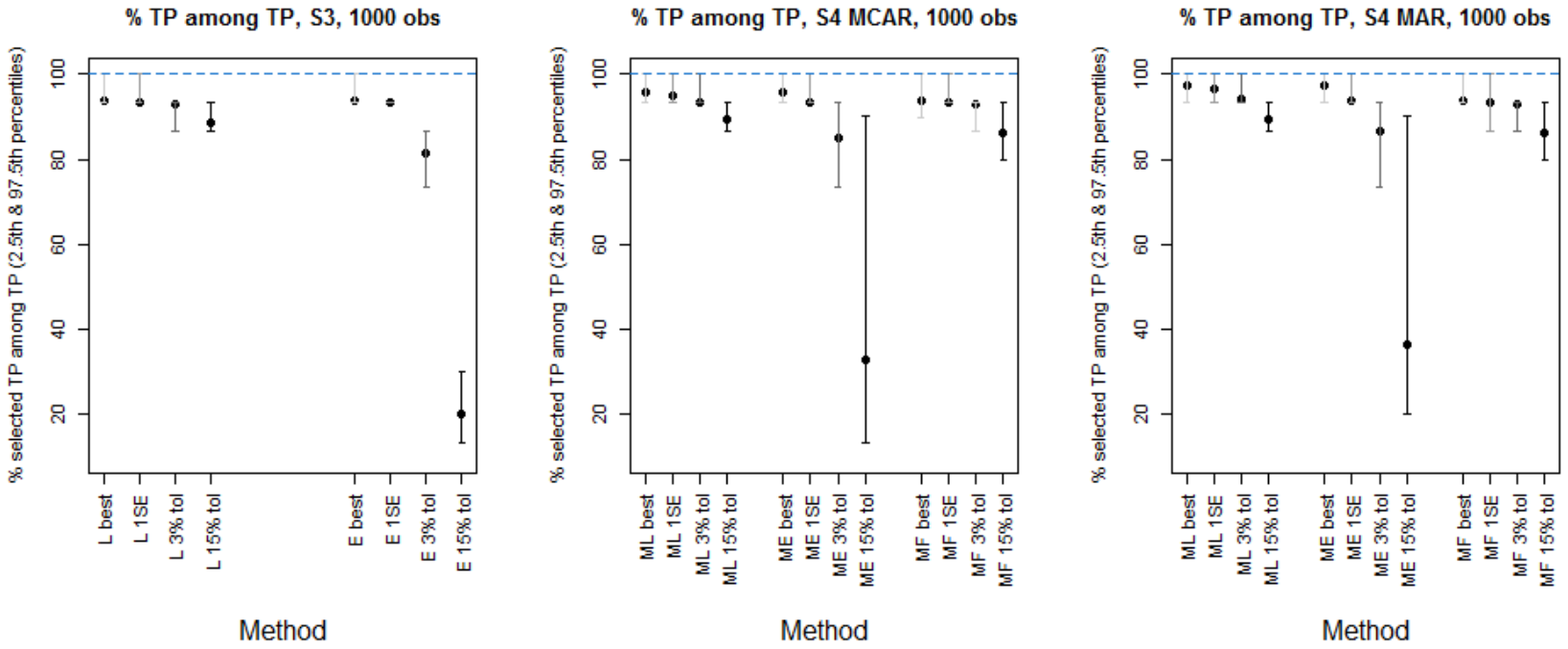


Figure A.16: Average percentage of **true predictors (TP) selected among the actual TP** (SEN) estimates with 2.5th and 97.5th percentiles from 3 methods run on 300 simulated **20-covariate** datasets with **1000 observations** for scenarios **S3** (assumption of moderation, without missing data) and **S5** (assumption of moderation, missing data also in the outcome). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME) and MissForest-Lasso (MF). ML, ME and MF estimated percentages of TP selected among the actual TP variables are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S3 (first plot from the left), the Lasso (L) and the Elasticnet (E) estimates are shown.

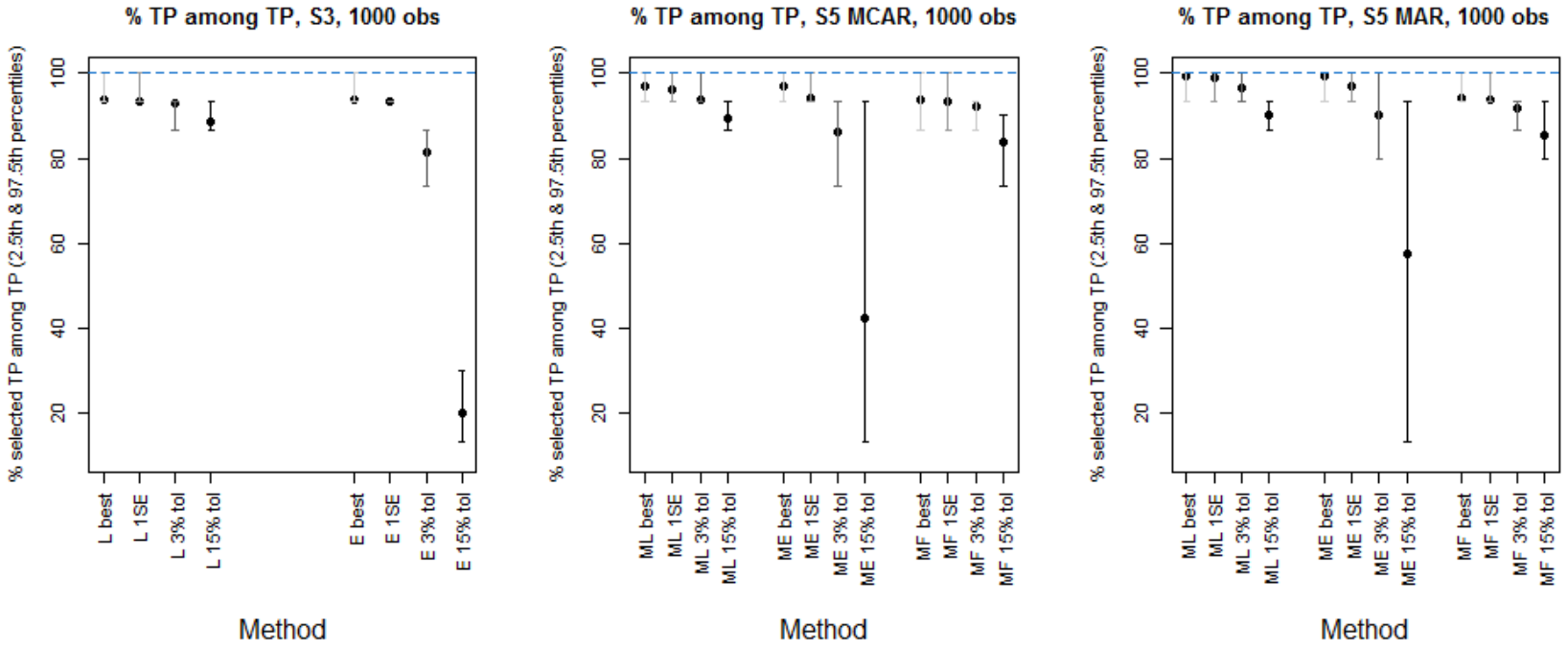


Figure A.17: Average percentage of **true predictors (TP) among the selected variables** (PPV) estimates with 2.5th and 97.5th percentiles from 3 methods run on 300 simulated **20-covariate** datasets with **1000 observations** for scenarios **S1** (without missing data) and **S2** (with missing data). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME) and MissForest-Lasso (MF). ML, ME and MF estimated percentages of TP among the selected variables are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S1 (first plot from the left), the Lasso (L) and the Elasticnet (E) estimates are shown.

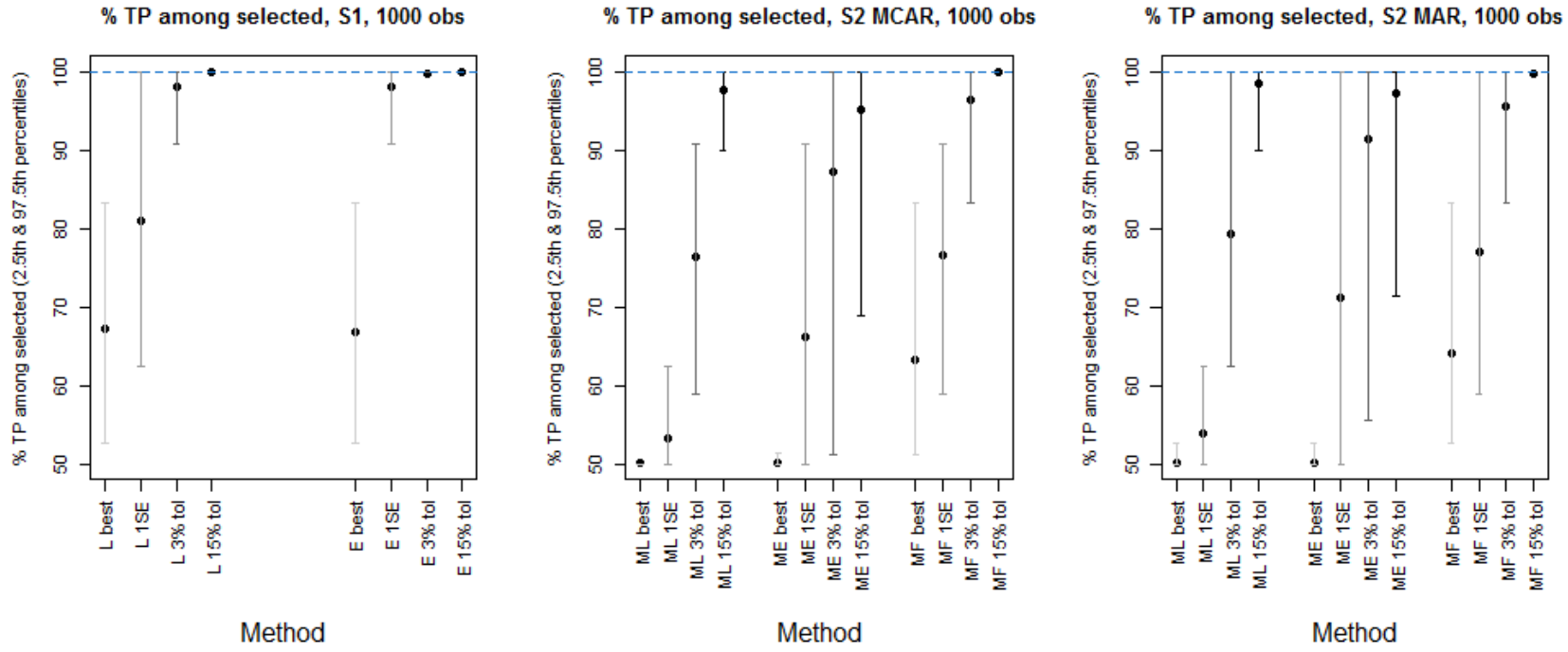


Figure A.18: Average percentage of **true predictors (TP) among the selected variables** (PPV) estimates with 2.5th and 97.5th percentiles from 3 methods run on 300 simulated **20-covariate** datasets with **1000 observations** for scenarios **S3** (assumption of moderation, without missing data) and **S4** (assumption of moderation, with missing data). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME) and MissForest-Lasso (MF). ML, ME and MF estimated percentages of TP among the selected variables are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S3 (first plot from the left), the Lasso (L) and the Elasticnet (E) estimates are shown.

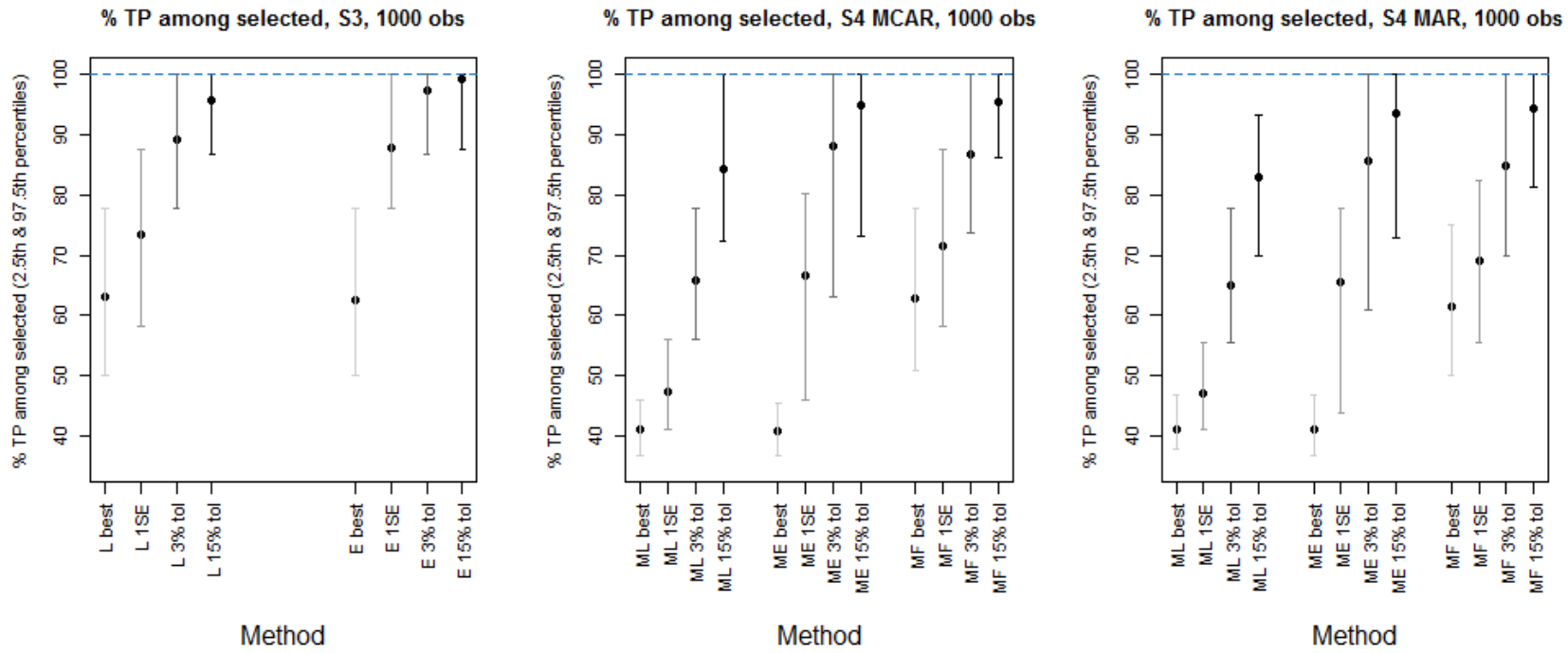


Figure A.19: Average percentage of **true predictors (TP) among the selected variables** (PPV) estimates with 2.5th and 97.5th percentiles from 3 methods run on 300 simulated **20-covariate** datasets with **1000 observations** for scenarios **S3** (assumption of moderation, without missing data) and **S5** (assumption of moderation, missing data also in the outcome). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME) and MissForest-Lasso (MF). ML, ME and MF estimated percentages of TP among the selected variables are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S3 (first plot from the left), the Lasso (L) and the Elasticnet (E) estimates are shown.

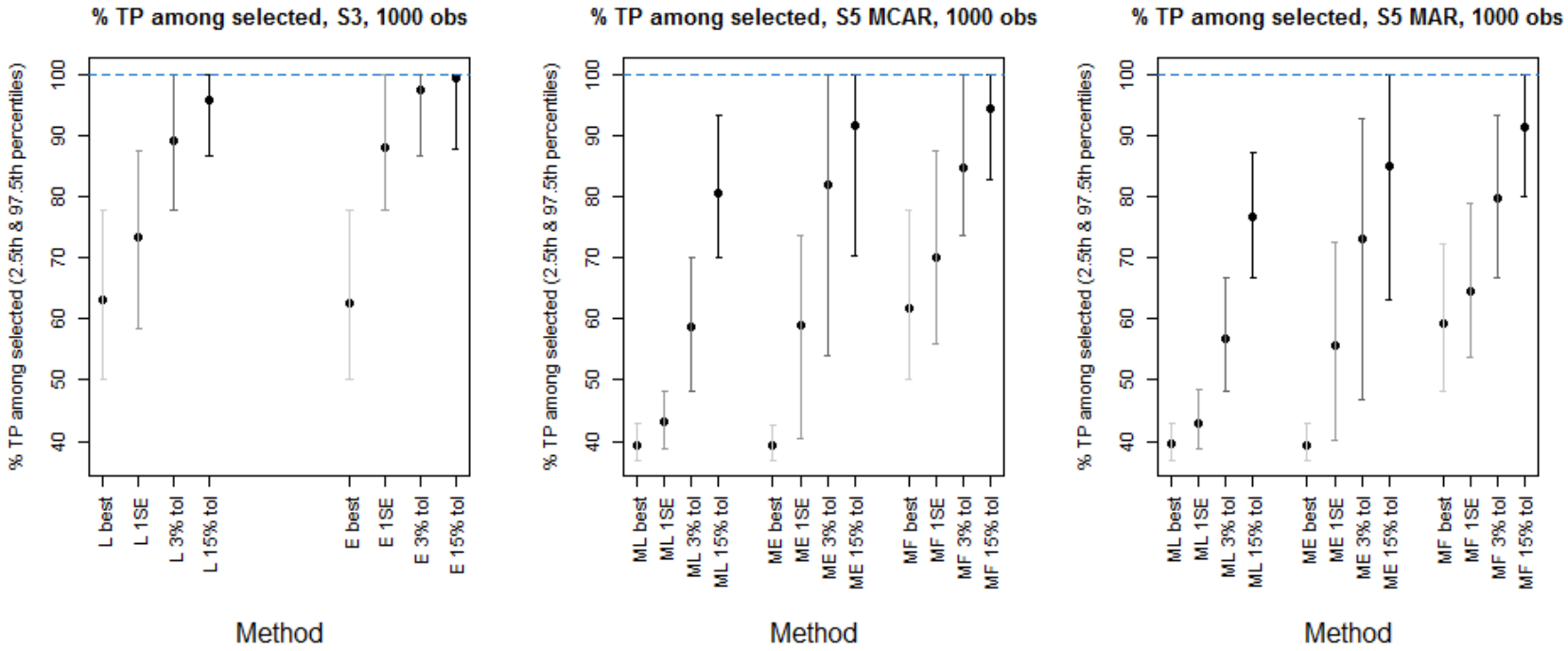


Figure A.20: Estimated percentage of **correct (true) models** (simultaneously with respect to all predictors) found by 4 methods run on 300 simulated **20-covariate** datasets with **1000 observations** for scenarios **S1** (without missing data) and **S2** (with missing data). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME), MissForest-Lasso (MF) and MissForest-Random Forests (MR). ML, ME and MF estimated percentages of selected true models are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S1 (first plot from the left), the Lasso (L), the Elasticnet (E) and the Random Forest (RF) estimates are shown. For the models RF and MR it is assumed that the true model is returned when the top 10 important variables are true predictors.

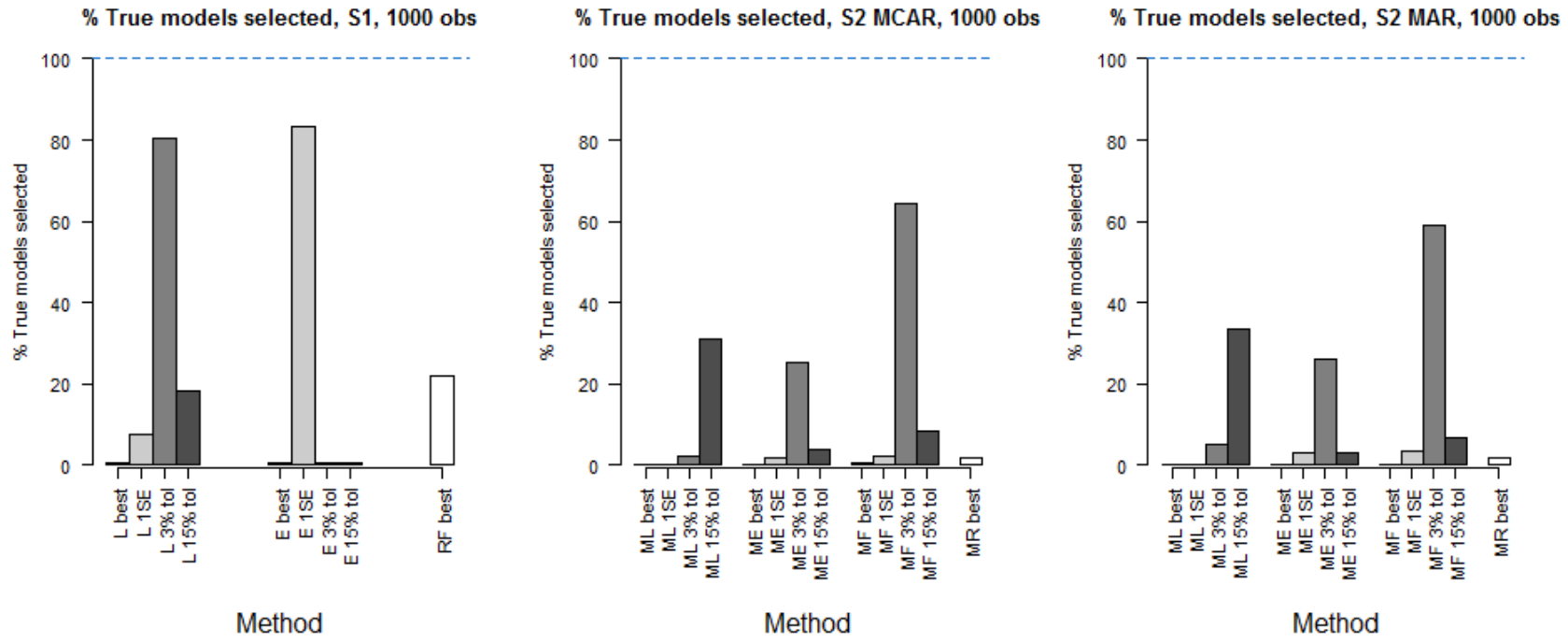


Figure A.21: Estimated percentage of **almost correct models** (only one variable off) found by 4 methods run on 300 simulated **20-covariate** datasets with **1000 observations** for scenarios **S1** (without missing data) and **S2** (with missing data). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME), MissForest-Lasso (MF) and MissForest-Random Forests (MR). ML, ME and MF estimated percentages of selected true models are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S1 (first plot from the left), the Lasso (L), the Elasticnet (E) and the Random Forests (RF) estimates are shown. For the models RF and MR, only the best model is computed through bootstrap tuning of the parameter given by the number of variables chosen randomly at each split to build the trees and it is assumed that the true model is returned when the top 10 important variables are the true predictors.

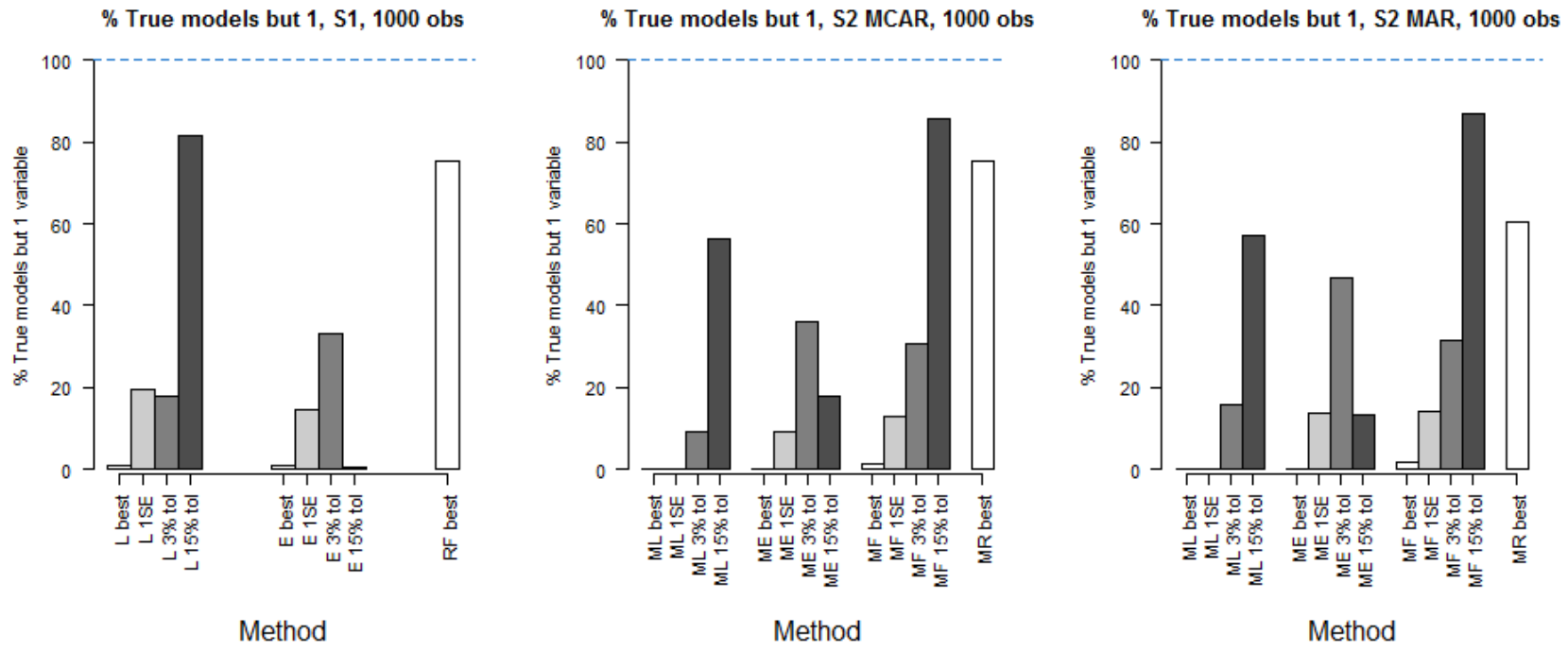


Figure A.22: Estimated percentage of **almost correct models** (only one variable off) found by 4 methods run on 300 simulated **20-covariate** datasets with **1000 observations** for scenarios **S3** (assumption of moderation, without missing data) and **S4** (assumption of moderation, with missing data). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME), MissForest-Lasso (MF) and MissForest-Random Forests (MR). ML, ME and MF estimated percentages of selected true models are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S3 (first plot from the left), the Lasso (L), the Elasticnet (E) and the Random Forests (RF) estimates are shown. For the models RF and MR, only the best model is computed through bootstrap tuning of the parameter given by the number of variables chosen randomly at each split to build the trees and it is assumed that the true model is returned when the top 10 important variables are the true predictors.

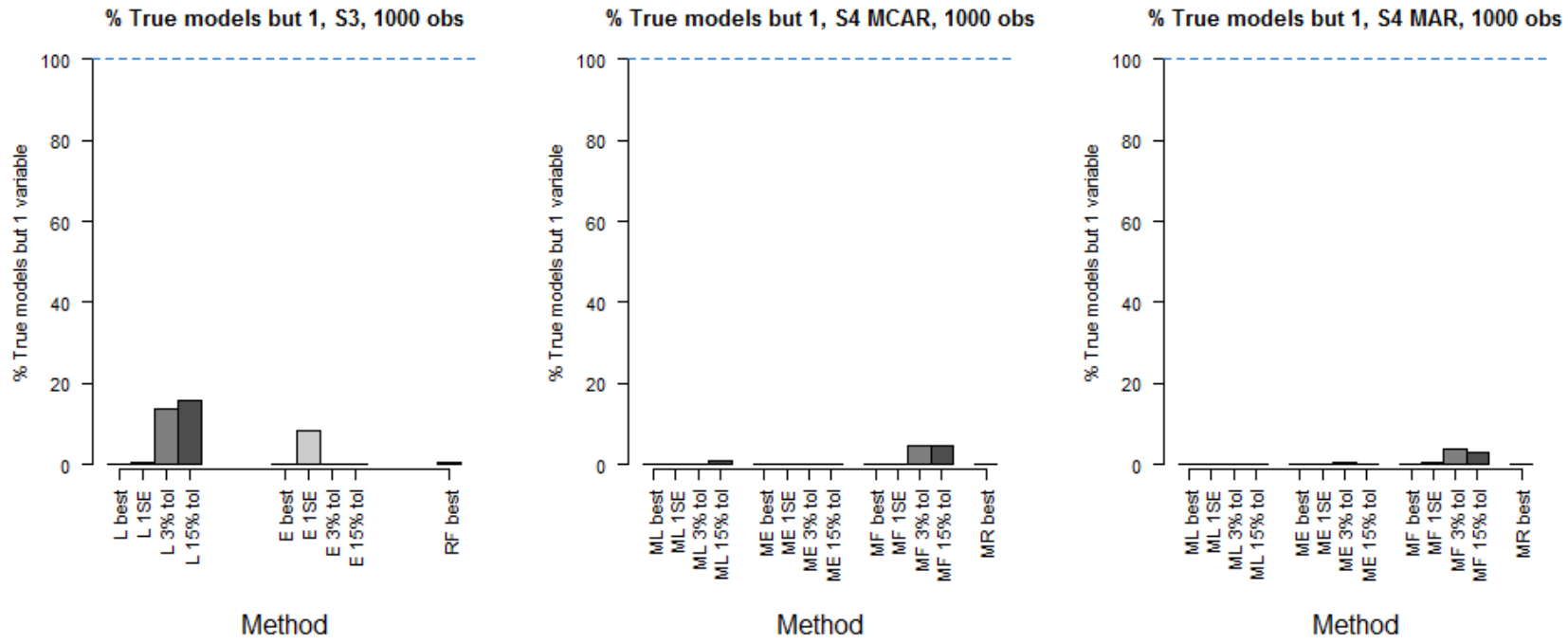


Figure A.23: Estimated percentage of **almost correct models** (only one variable off) found by 4 methods run on 300 simulated **20-covariate** datasets with **1000 observations** for scenarios **S3** (assumption of moderation, without missing data) and **S5** (assumption of moderation, missing data also in the outcome). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME), MissForest-Lasso (MF) and MissForest-Random Forests (MR). ML, ME and MF estimated percentages of selected true models are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For S3 (first plot from the left), the Lasso (L), the Elasticnet (E) and the Random Forests (RF) estimates are shown. For the models RF and MR, only the best model is computed through bootstrap tuning of the parameter given by the number of variables chosen randomly at each split to build the trees and it is assumed that the true model is returned when the top 10 important variables are the true predictors.

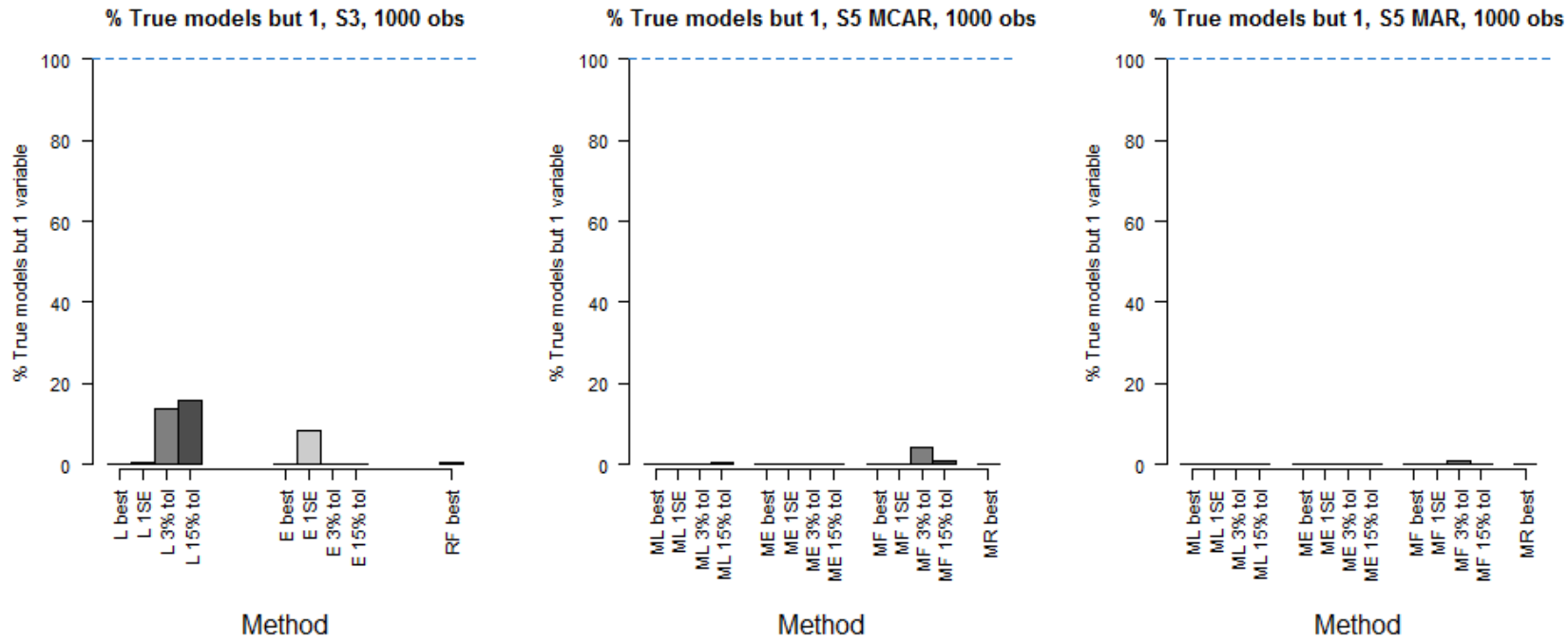


Figure A.24: Comparison of **variable inclusion frequency** by 3 methods run on 300 simulated **20-covariate** datasets with **1000 observations** for scenario **S1** (no assumption of moderation, complete data). The methods are: Lasso, Elasticnet and Random Forest. Lasso and Elasticnet variable inclusion frequencies are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For Random Forests only the best model is computed through bootstrap tuning of the parameter given by the number of variables chosen randomly at each split to build the trees and a variable is considered included in the model when its importance is among the top 10 variable importances.

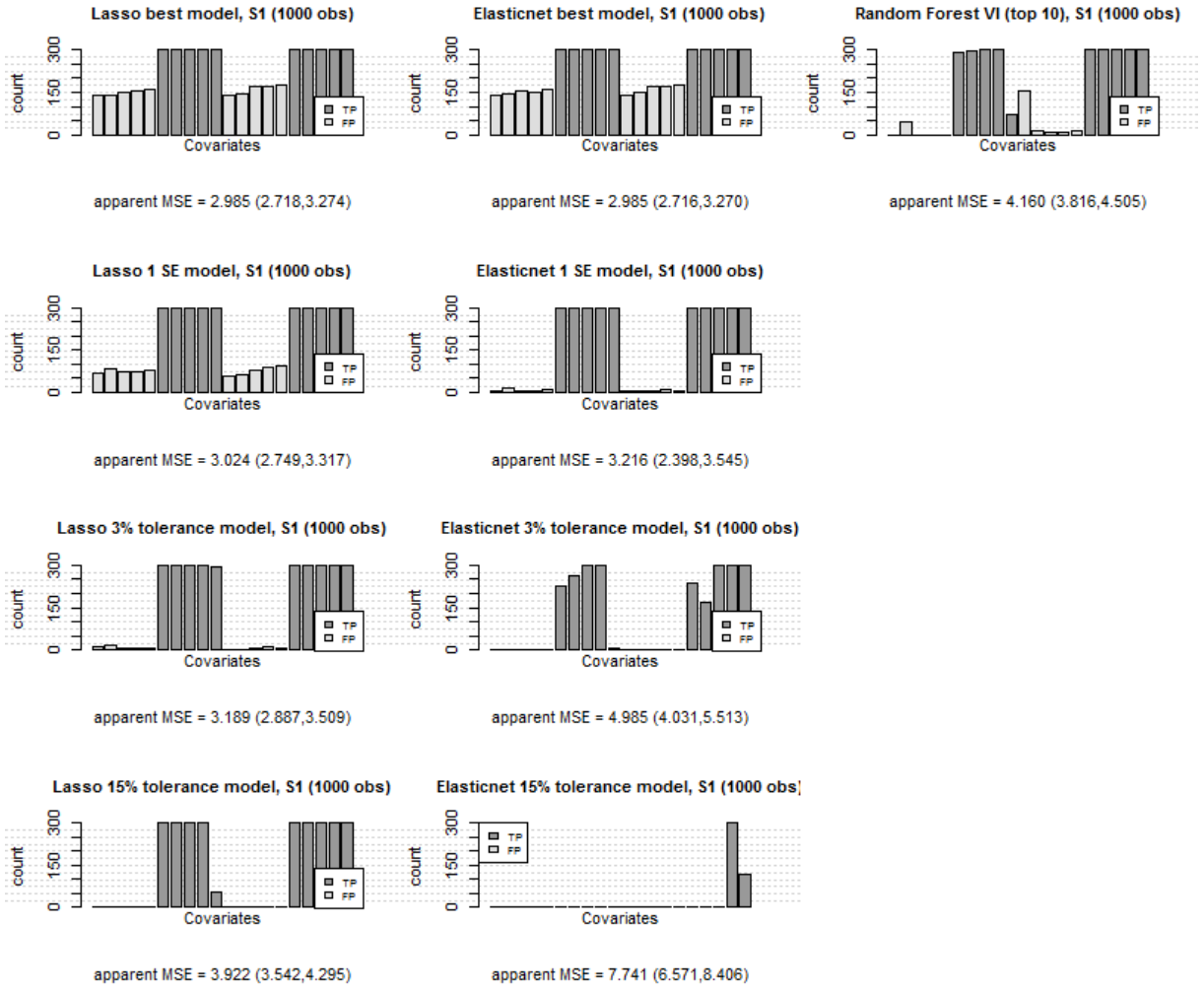


Figure A.25: Comparison of **variable inclusion frequency** by 3 methods run on 300 simulated **20-covariate** datasets with **1000 observations** for scenario **S2** with **MCAR** data (no assumption of moderation, complete outcome). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME), MissForest-Lasso (MF) and MissForest-Random Forests (MR). ML, ME and MF variable inclusion frequencies are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For MR only the best model is computed through bootstrap tuning of the parameter given by the number of variables chosen randomly at each split to build the trees and a variable is considered included in the model when its importance is among the top 10 variable importances.

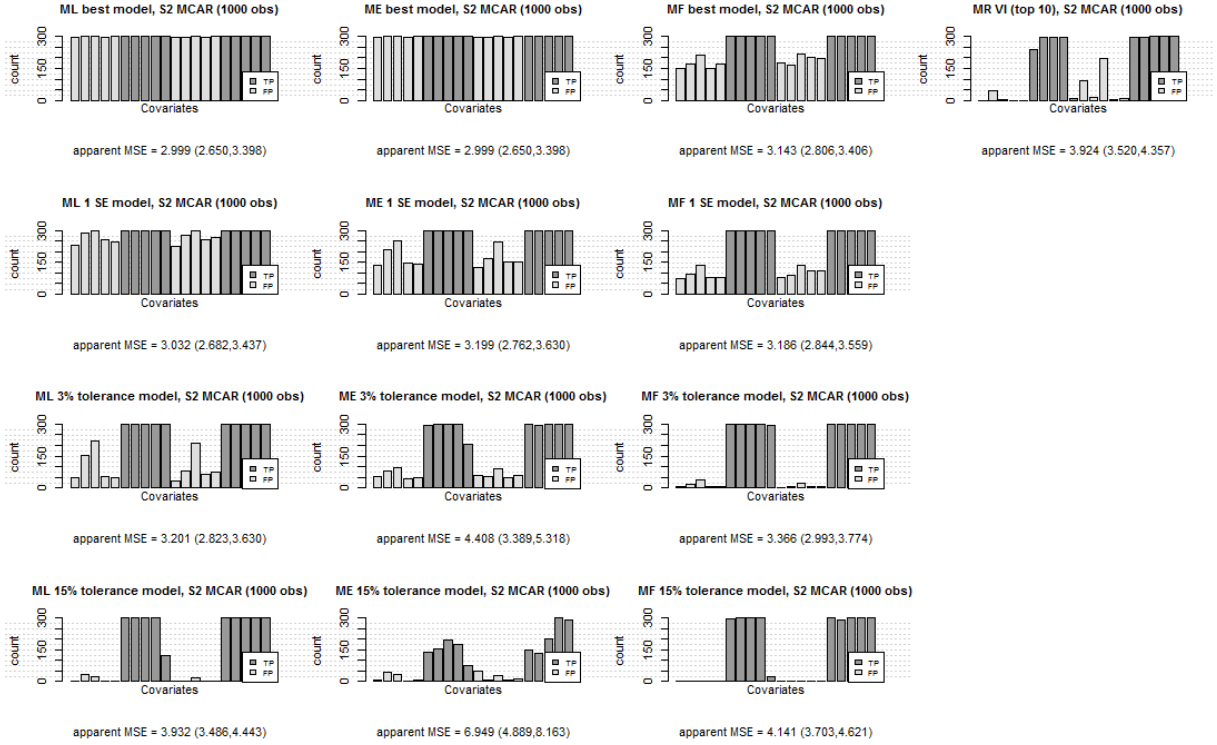


Figure A.26: Comparison of **variable inclusion frequency** by 3 methods run on 300 simulated **20-covariate** datasets with **1000 observations** for scenario **S2** with **MAR** data (no assumption of moderation, complete outcome). MICE-Lasso (ML), MICE-Elasticnet (ME), MissForest-Lasso (MF) and MissForest-Random Forests (MR). ML, ME and MF variable inclusion frequencies are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For MR only the best model is computed through bootstrap tuning of the parameter given by the number of variables chosen randomly at each split to build the trees and a variable is considered included in the model when its importance is among the top 10 variable importances.

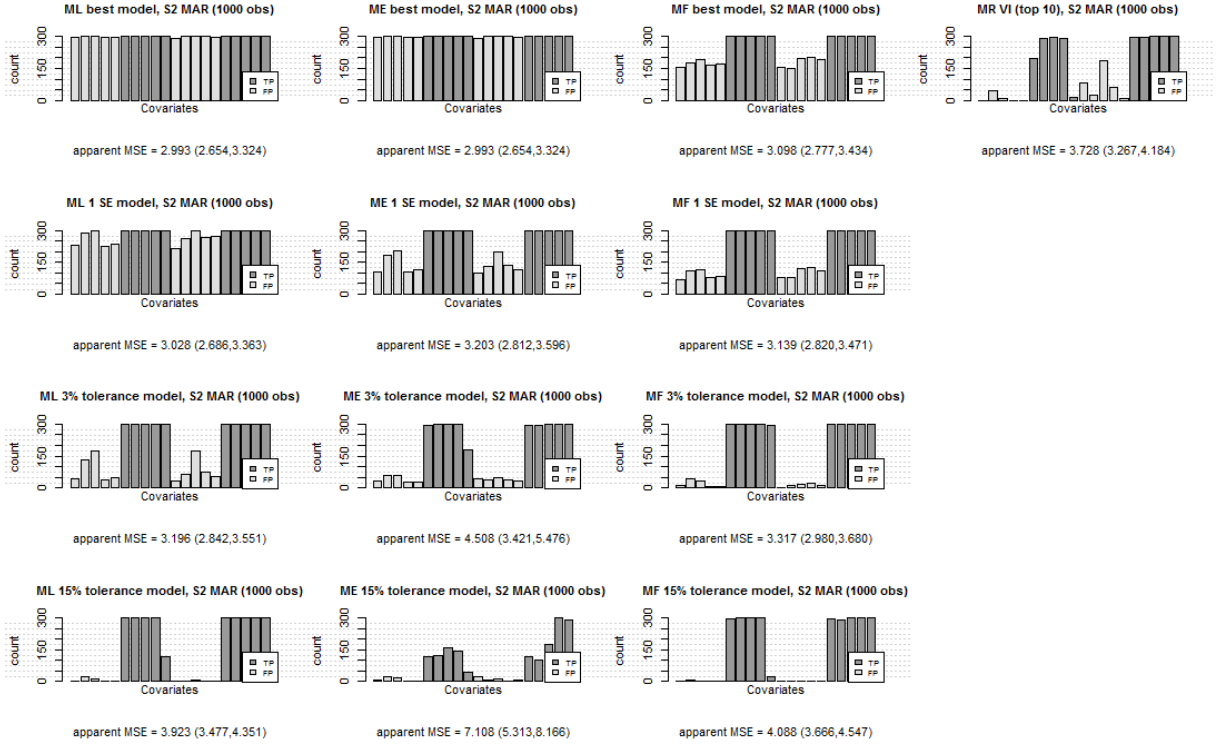


Figure A.27: Comparison of **variable inclusion frequency** by 3 methods run on 300 simulated **20-covariate** datasets with **1000 observations** for scenario **S3** (assumption of moderation, complete data). The methods are: Lasso, Elasticnet and Random Forest. Lasso and Elasticnet variable inclusion frequencies are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For Random Forests only the best model is computed through bootstrap tuning of the parameter given by the number of variables chosen randomly at each split to build the trees and a variable is considered included in the model when its importance is among the top 10 variable importances.

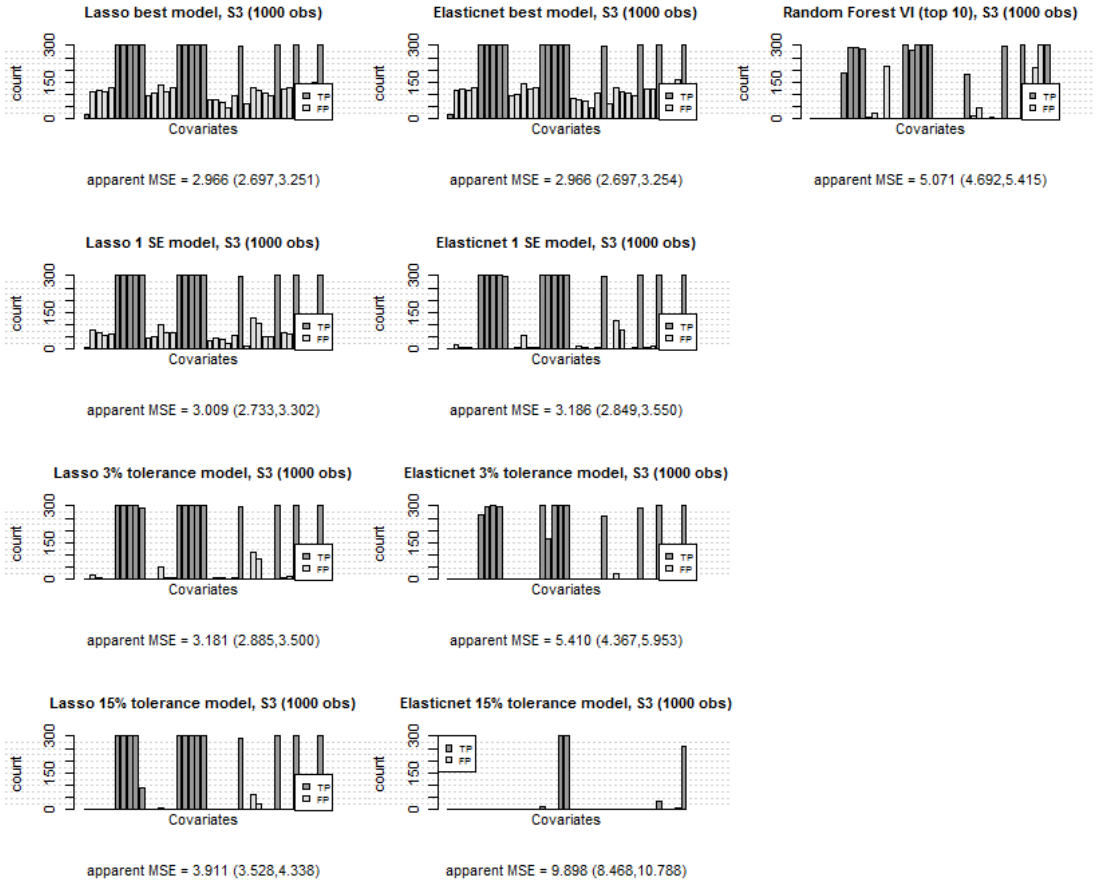


Figure A.28: Comparison of **variable inclusion frequency** by 3 methods run on 300 simulated **20-covariate** datasets with **1000 observations** for scenario **S4** with **MCAR** data (assumption of moderation, complete outcome). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME), MissForest-Lasso (MF) and MissForest-Random Forests (MR). ML, ME and MF variable inclusion frequencies are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For MR only the best model is computed through bootstrap tuning of the parameter given by the number of variables chosen randomly at each split to build the trees and a variable is considered included in the model when its importance is among the top 10 variable importances.

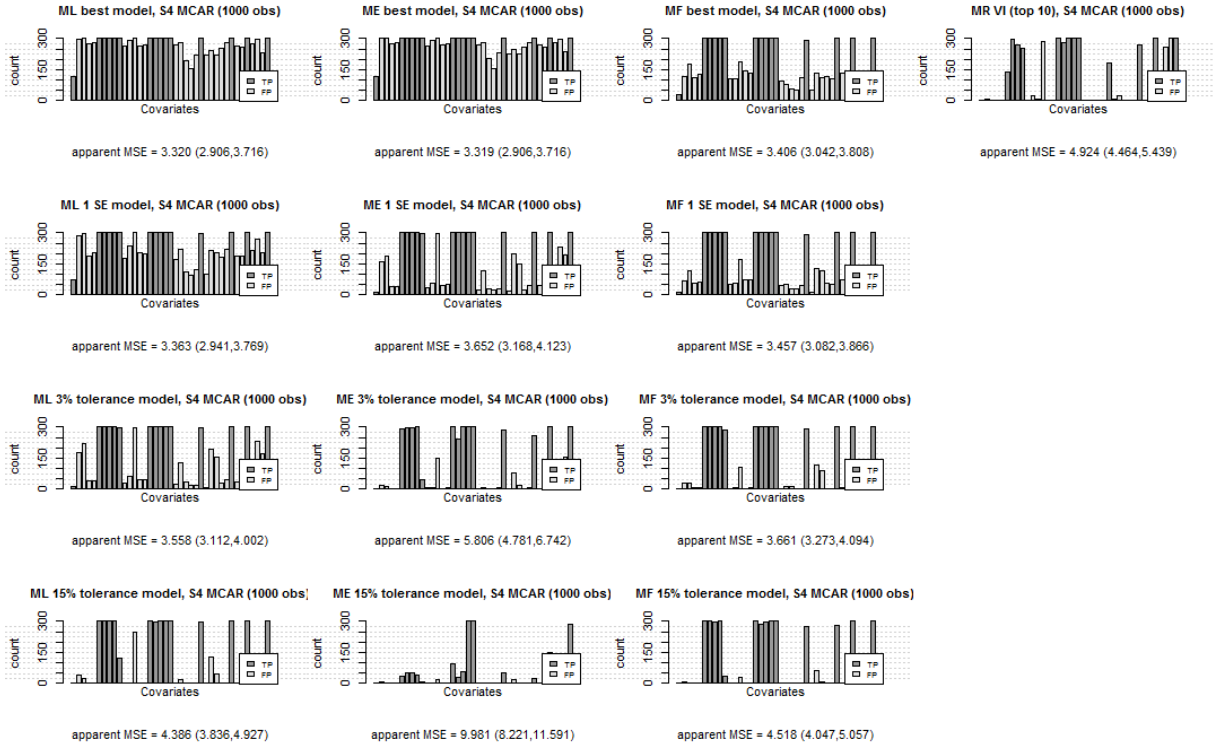


Figure A.29: Comparison of **variable inclusion frequency** by 3 methods run on 300 simulated **20-covariate** datasets with **1000 observations** for scenario **S4** with **MAR** data (assumption of moderation, complete outcome). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME), MissForest-Lasso (MF) and MissForest-Random Forests (MR). ML, ME and MF variable inclusion frequencies are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For MR only the best model is computed through bootstrap tuning of the parameter given by the number of variables chosen randomly at each split to build the trees and a variable is considered included in the model when its importance is among the top 10 variable importances.

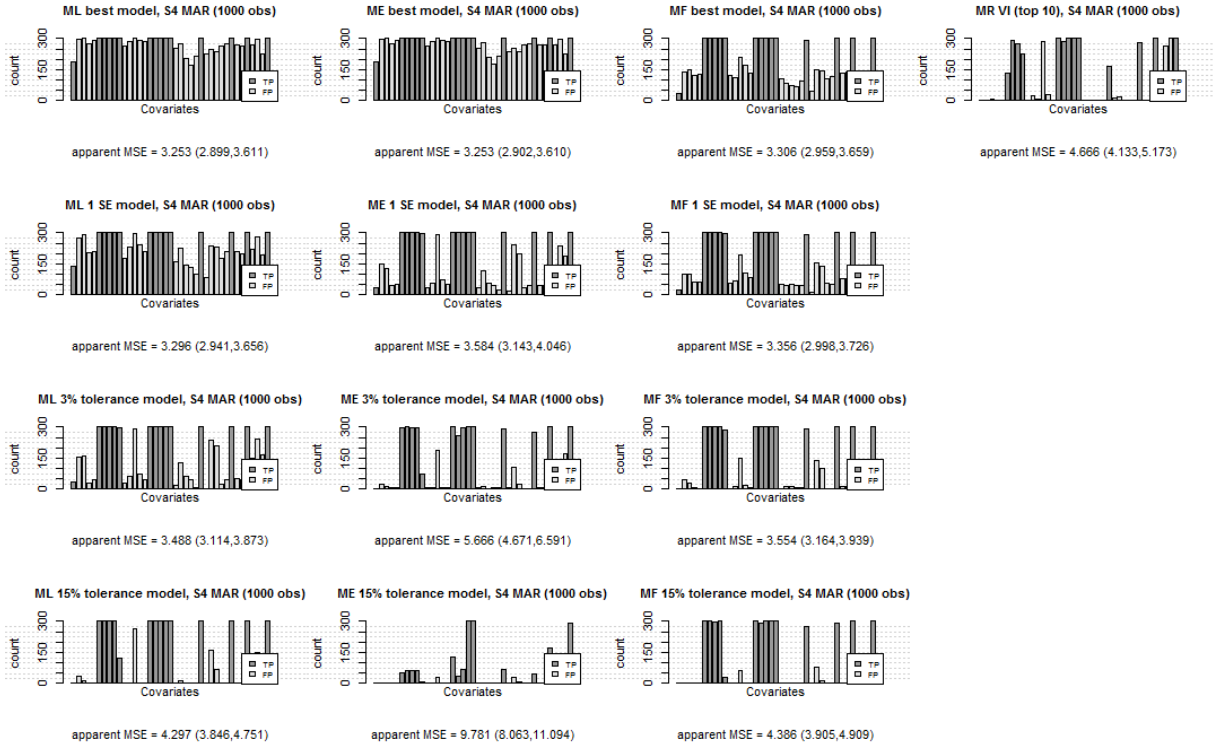


Figure A.30: Comparison of **variable inclusion frequency** by 3 methods run on 300 simulated **20-covariate** datasets with **1000 observations** for scenario **S5** with **MCAR** data (assumption of moderation, missing data also in the outcome). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME), MissForest-Lasso (MF) and MissForest-Random Forests (MR). ML, ME and MF variable inclusion frequencies are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For MR only the best model is computed through bootstrap tuning of the parameter given by the number of variables chosen randomly at each split to build the trees and a variable is considered included in the model when its importance is among the top 10 variable importances.

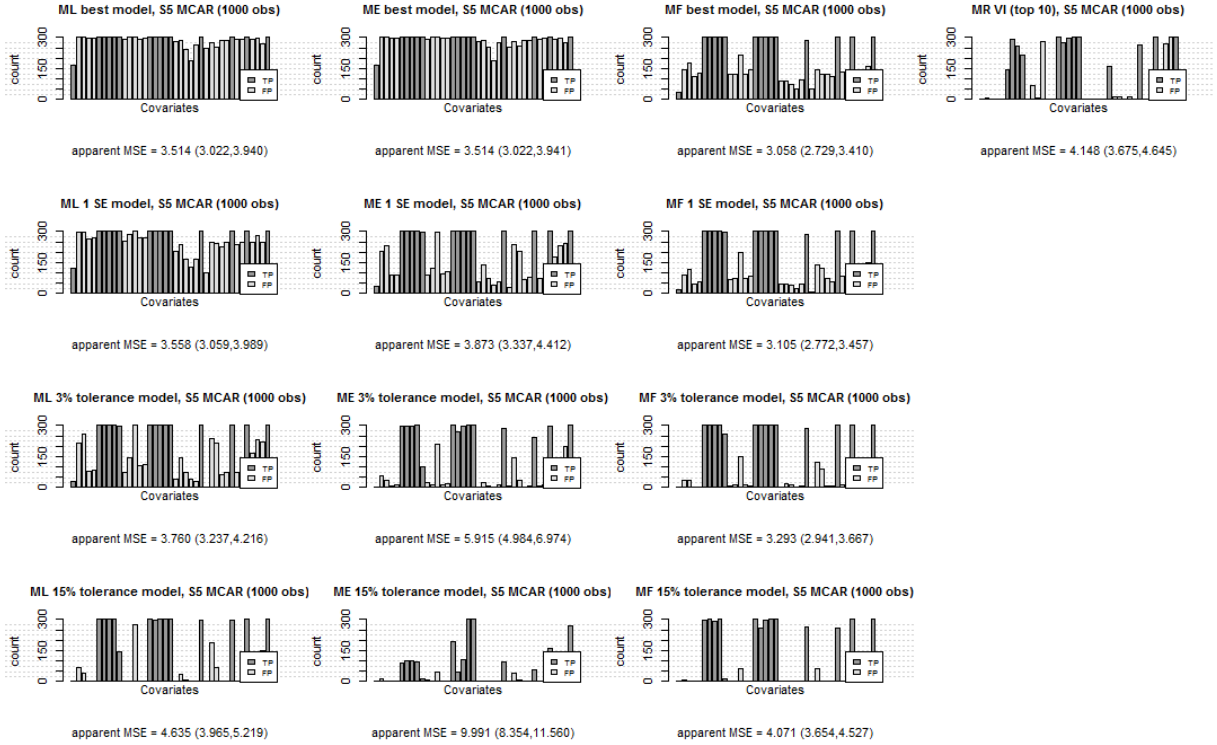


Figure A.31: Comparison of **variable inclusion frequency** by 3 methods run on 300 simulated **20-covariate** datasets with **1000 observations** for scenario **S5** with **MAR** data (assumption of moderation, missing data also in the outcome). The methods are: MICE-Lasso (ML), MICE-Elasticnet (ME), MissForest-Lasso (MF) and MissForest-Random Forests (MR). ML, ME and MF variable inclusion frequencies are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For MR only the best model is computed through bootstrap tuning of the parameter given by the number of variables chosen randomly at each split to build the trees and a variable is considered included in the model when its importance is among the top 10 variable importances.

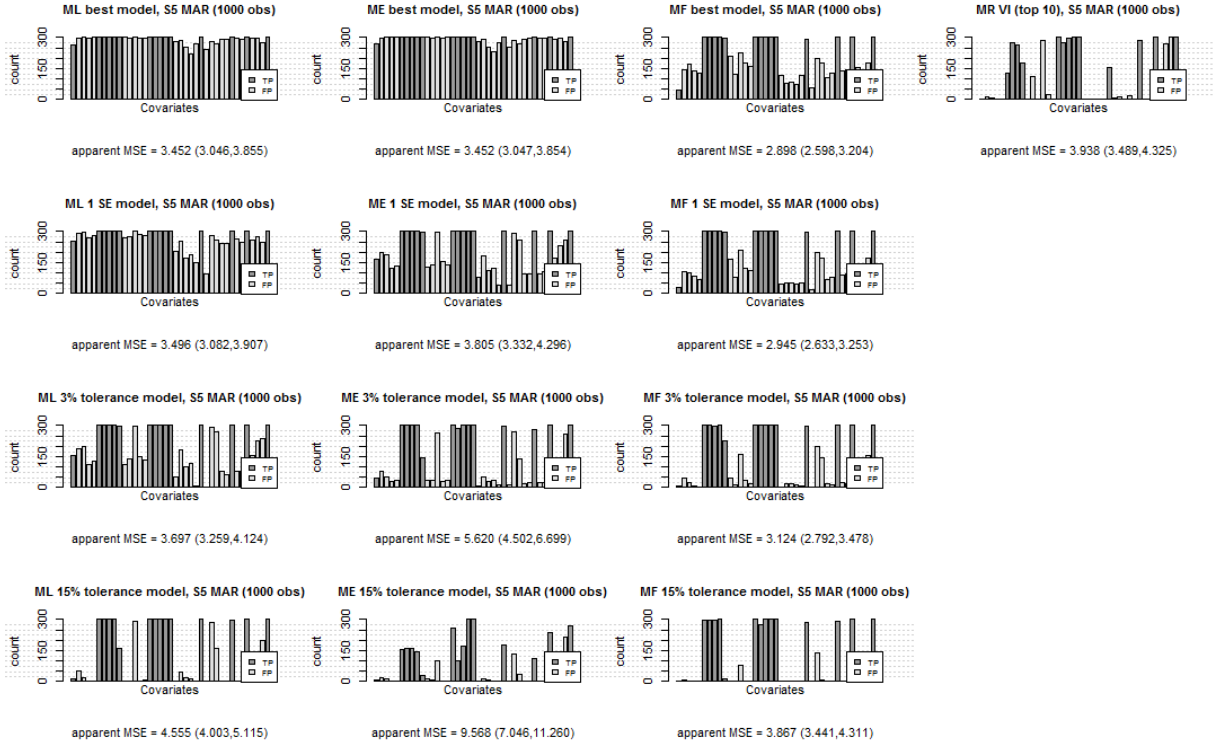


Figure A.32: Comparison of **variable inclusion frequency** by 3 methods run on 300 simulated **100-covariate** datasets with **500 observations** and between-covariate **correlation of 0.2** for scenario **S3** (assumption of moderation, complete data). The methods are: Lasso, Elasticnet and Conditional RF. Lasso and Elasticnet variable inclusion frequencies are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For Conditional Random Forests (RF) only the best model is computed through bootstrap tuning of the parameter given by the number of variables chosen randomly at each split to build the trees and a variable is considered included in the model when its importance is among the top 15 variable importances. Note that **Conditional RF always had exactly the TPs as the 15 most important variables**.

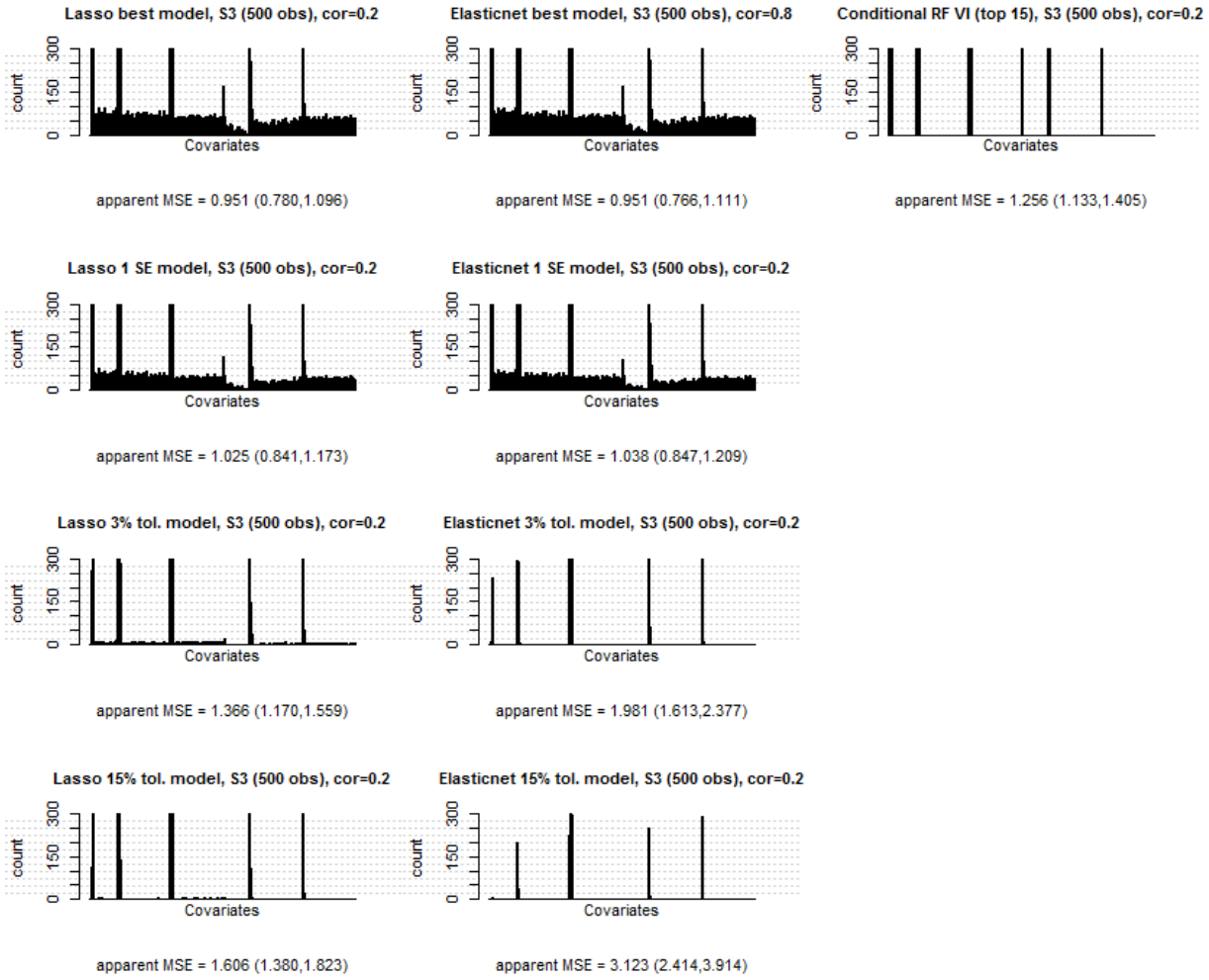


Figure A.33: Comparison of **variable inclusion frequency** by 3 methods run on 300 simulated **100-covariate** datasets with **500 observations** and between-covariate **correlation of 0.8** for scenario **S3** (assumption of moderation, complete data). The methods are: Lasso, Elasticnet and Conditional RF. Lasso and Elasticnet variable inclusion frequencies are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For Conditional Random Forests (RF) only the best model is computed through bootstrap tuning of the parameter given by the number of variables chosen randomly at each split to build the trees and a variable is considered included in the model when its importance is among the top 15 variable importances. Note that **Conditional RF always had exactly the TPs as the 15 most important variables**.

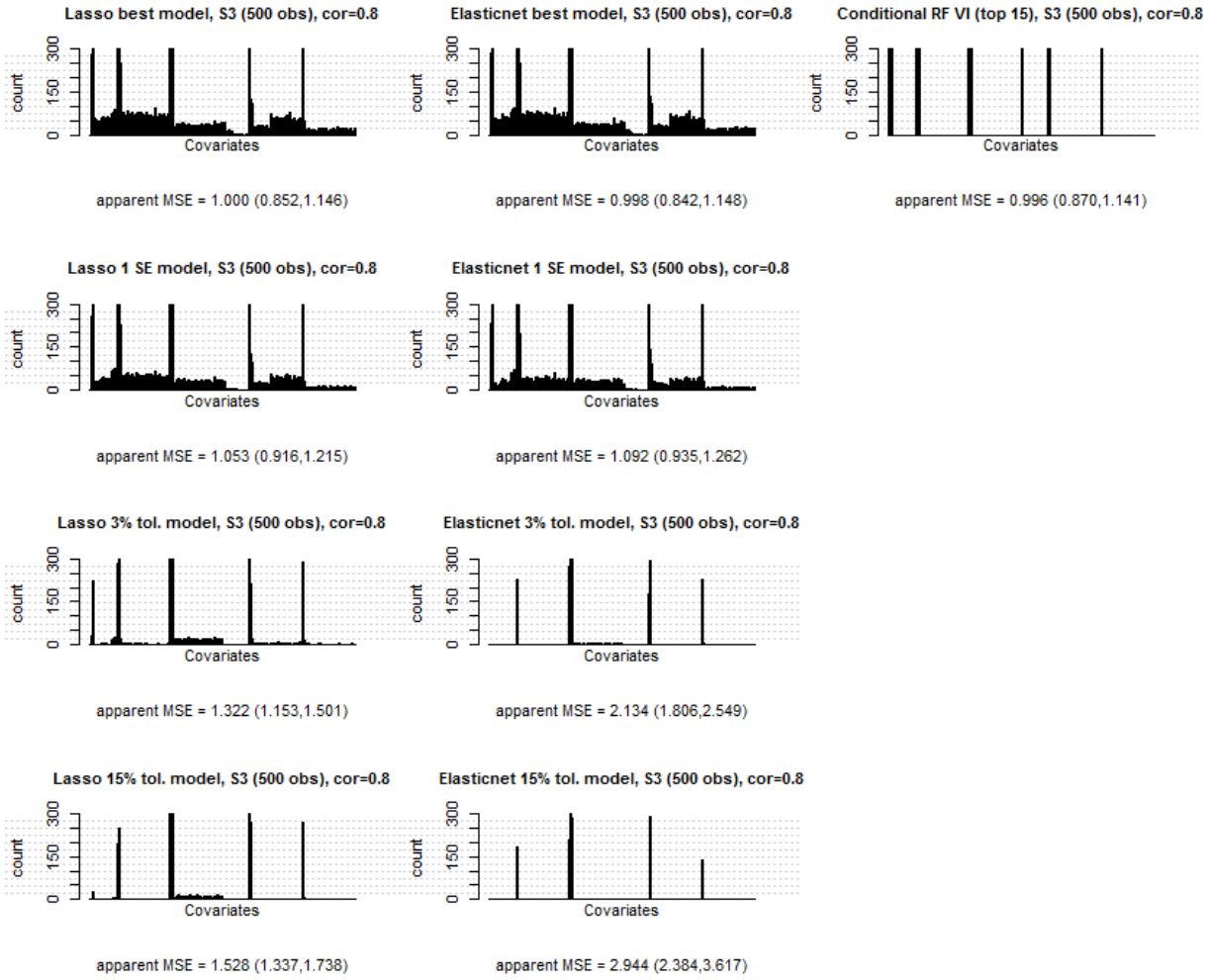


Figure A.34: Comparison of **variable inclusion frequency** by 3 methods run on 300 simulated **100-covariate** datasets with **500 observations** and between-covariate **correlation of 0.2** for scenario **S5** with **MCAR** data (assumption of moderation, missing data also in the outcome). The methods are: Lasso, Elasticnet and Conditional RF. Lasso and Elasticnet variable inclusion frequencies are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For Conditional Random Forests (RF) only the best model is computed through bootstrap tuning of the parameter given by the number of variables chosen randomly at each split to build the trees and a variable is considered included in the model when its importance is among the top 15 variable importances. Note that **Conditional RF always had exactly the TPs as the 15 most important variables**.

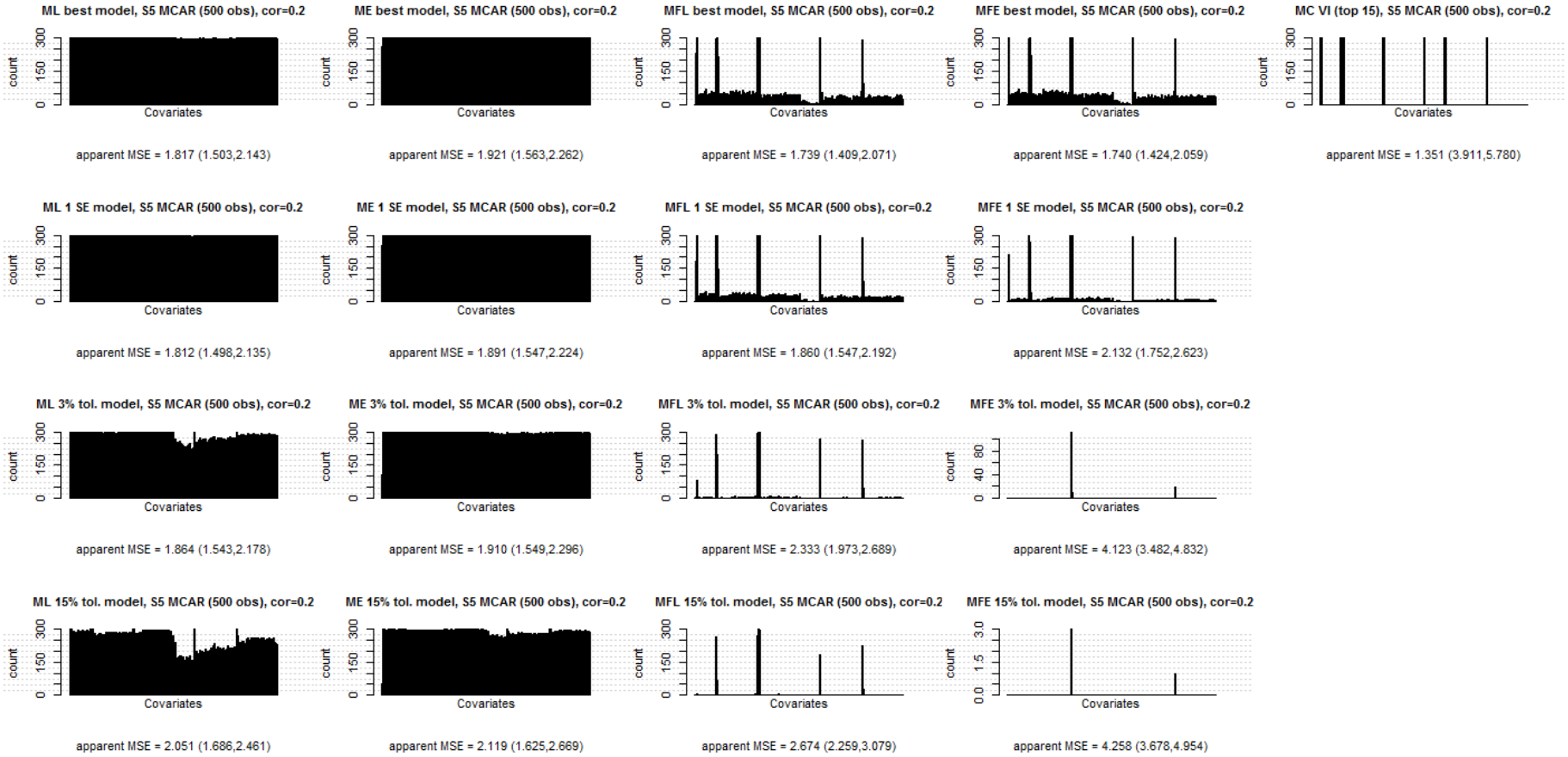


Figure A.35: Comparison of **variable inclusion frequency** by 3 methods run on 300 simulated **100-covariate** datasets with **500 observations** and between-covariate **correlation of 0.8** for scenario **S5** with **MCAR** data (assumption of moderation, missing data also in the outcome). The methods are: Lasso, Elasticnet and Conditional RF. Lasso and Elasticnet variable inclusion frequencies are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For Conditional Random Forests (RF) only the best model is computed through bootstrap tuning of the parameter given by the number of variables chosen randomly at each split to build the trees and a variable is considered included in the model when its importance is among the top 15 variable importances. Note that **Conditional RF always had exactly the TPs as the 15 most important variables**.

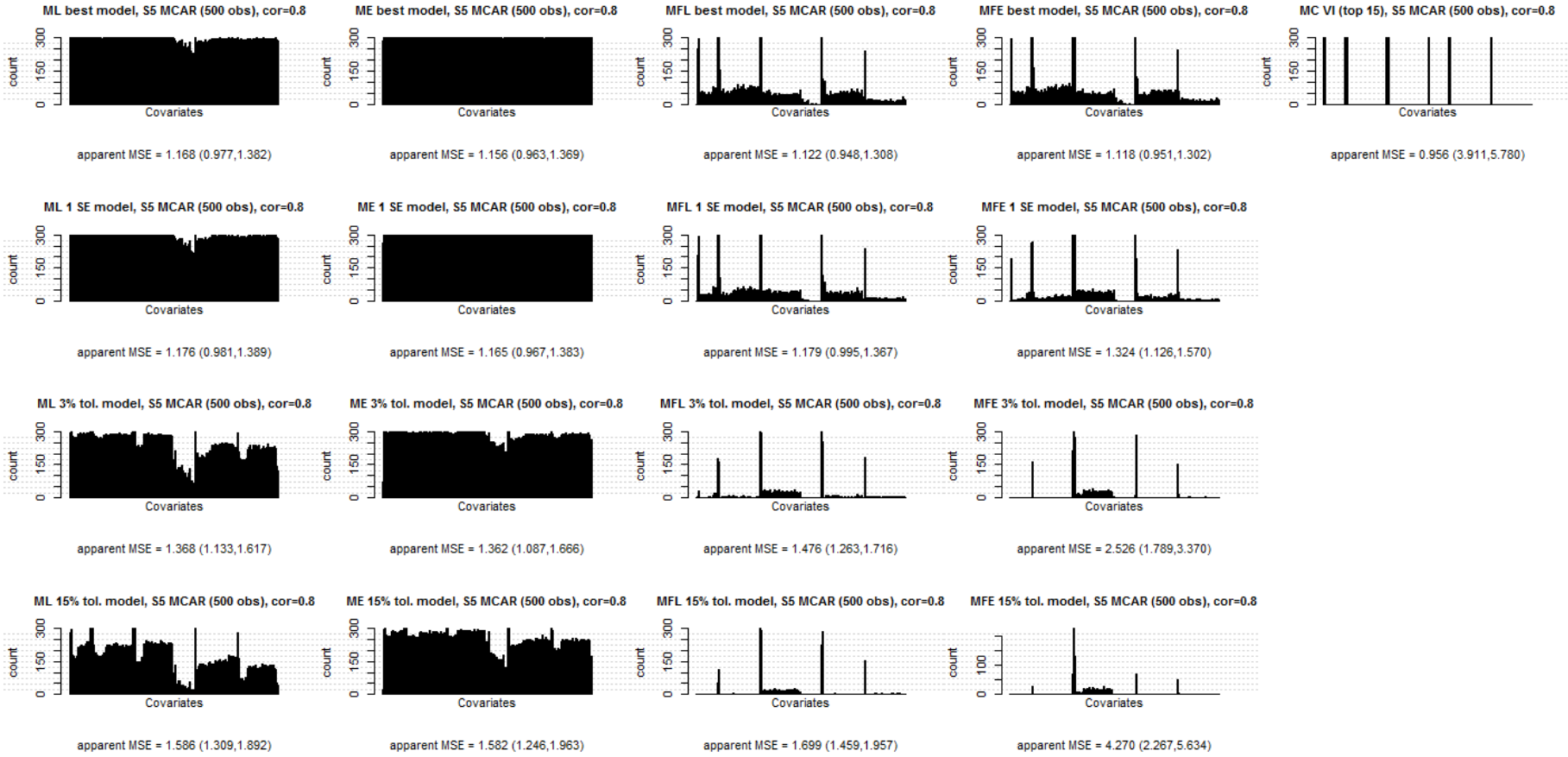


Figure A.36: Comparison of **variable inclusion frequency** by 3 methods run on 300 simulated **100-covariate** datasets with **500 observations** and between-covariate **correlation of 0.2** for scenario **S5** with **MAR** data (assumption of moderation, missing data also in the outcome). The methods are: Lasso, Elasticnet and Conditional RF. Lasso and Elasticnet variable inclusion frequencies are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For Conditional Random Forests (RF) only the best model is computed through bootstrap tuning of the parameter given by the number of variables chosen randomly at each split to build the trees and a variable is considered included in the model when its importance is among the top 15 variable importances. Note that **Conditional RF always had exactly the TPs as the 15 most important variables**.

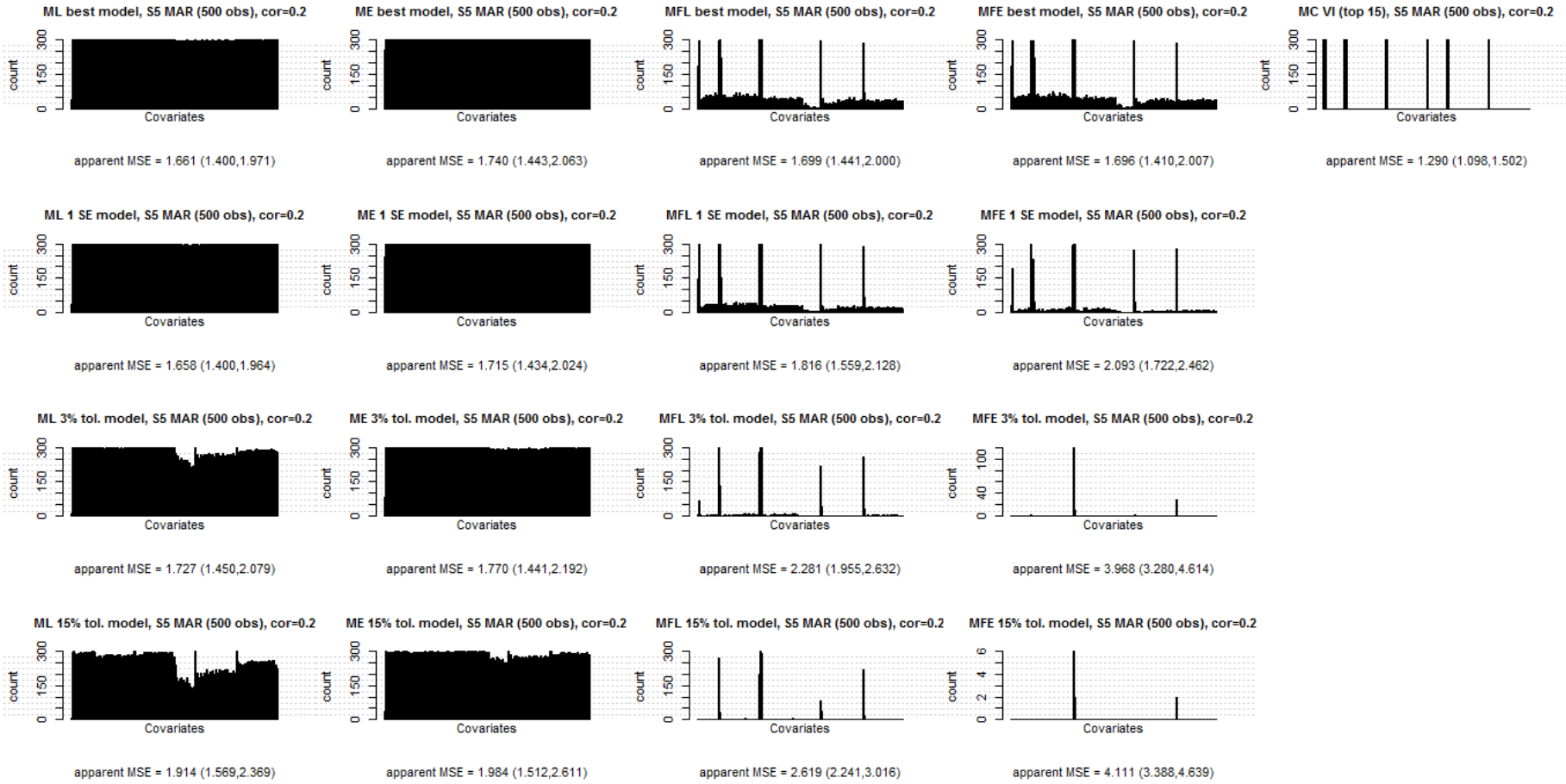
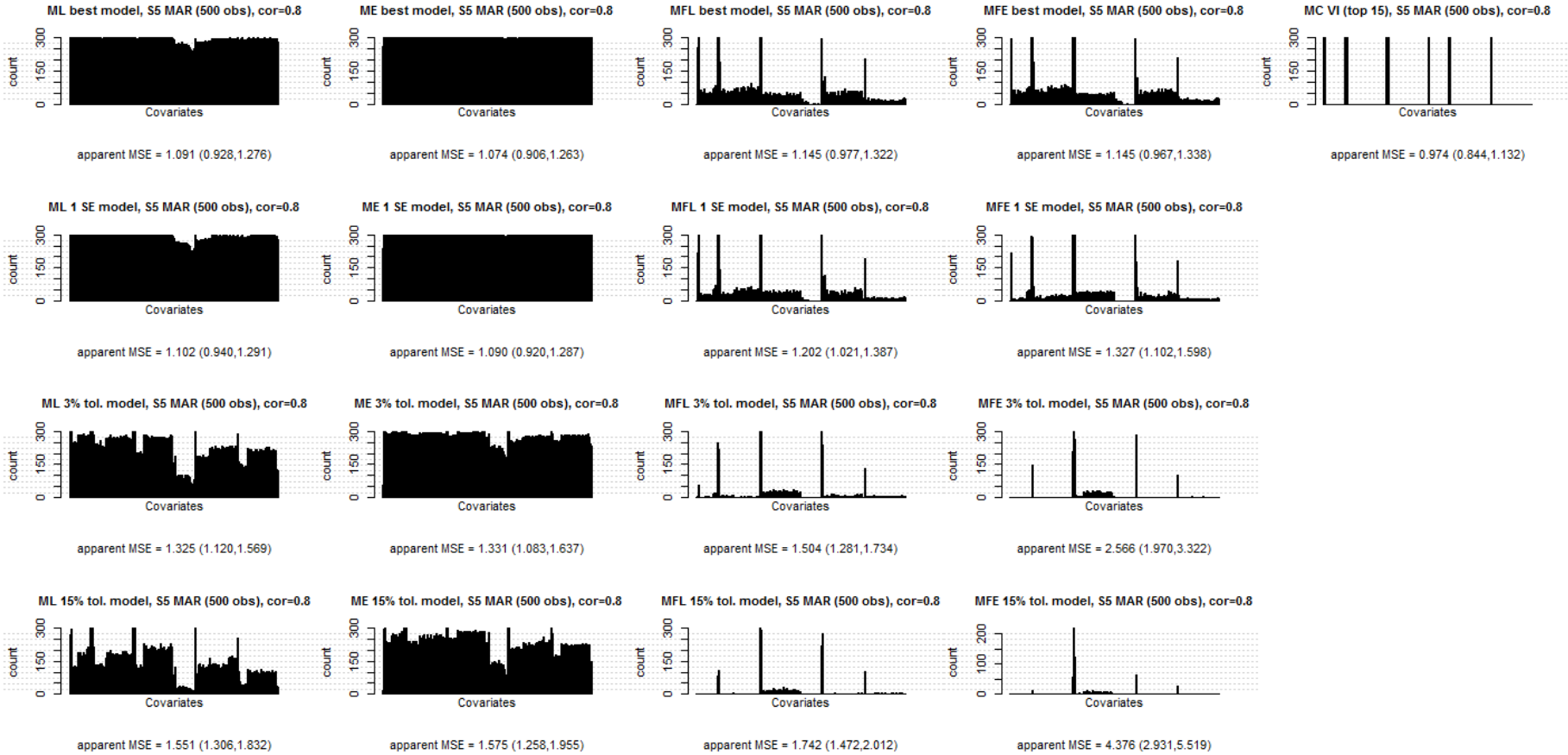


Figure A.37: Comparison of **variable inclusion frequency** by 3 methods run on 300 simulated **100-covariate** datasets with **500 observations** and between-covariate **correlation of 0.8** for scenario **S5** with **MAR** data (assumption of moderation, missing data also in the outcome). The methods are: Lasso, Elasticnet and Conditional RF. Lasso and Elasticnet variable inclusion frequencies are shown for the best $\lambda$ selection as well as for three tolerance models: one model corresponding to a $\lambda$ giving the MSE within 1 standard error (SE) of the minimum, the 2nd within 3% and the 3rd within 15%, through bootstrap tuning. For Conditional Random Forests (RF) only the best model is computed through bootstrap tuning of the parameter given by the number of variables chosen randomly at each split to build the trees and a variable is considered included in the model when its importance is among the top 15 variable importances. Note that **Conditional RF always had exactly the TPs as the 15 most important variables**.



## A.4 Selection of moderators

Table A.17: Comparison of average sensitivity (SEN), false positive rate (FPR) and positive predictive value (PPV) of selection for the predictors (P) and for the moderators (M) for the best $\lambda$ models in the simulation study. Average SEN, FPR and PPV are given in percentages with corresponding SD.

| | **Selection of predictors and moderators in the best $\lambda$ models** | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Scenario 3** | | **Scenario 4: complete outcome** | | | | **Scenario 5: incomplete outcome (20%)** | | | | **Scenario 6: interactions in imputation model** | | | |
| Data | Complete | | MCAR | | MAR | | MCAR | | MAR | | MCAR | | MAR | |
| **20-covariate** | N=250 | N=1000 | N=250 | N=1000 | N=250 | N=1000 | N=250 | N=1000 | N=250 | N=1000 | N=250 | N=1000 | N=250 | N=1000 |
| **Mice-Lasso** | | | | | | | | | | | | | | |
| SEN of P (SD) | 92.3 (3.3) | 100 (1.8) | 95.5 (3.3) | 95.9 (3.2) | 96 (3.5) | 97.4 (3.2) | 96.6 (3.4) | 97 (3.3) | 97.5 (3.2) | 99.2 (2.1) | 98.6 (2.7) | 97.5 (3.2) | 98.2 (3.2) | 97 (3.3) |
| SEN of M (SD) | 96.6 (8.6) | 99.8 (2.0) | 98.6 (3.2) | 100 (0) | 99.6 (3.2) | 100 (0) | 99.5 (3.5) | 100 (0) | 99.9 (1.4) | 100 (0) | 100 (0) | 100 (0) | 99.5 (3.5) | 99.9 (1.4) |
| FPR of P (SD) | 35.9 (11.9) | 51.3 (11.1) | 86.6 (8.8) | 86.3 (7.3) | 85.6 (9) | 87.3 (7.4) | 94.1 (5) | 93.6 (4.8) | 93.6 (5.3) | 94.9 (5) | 98.4 (3.7) | 97.4 (4.1) | 96.5 (5.8) | 96.7 (4) |
| FPR of M (SD) | 33.4 (14.4) | 33.9 (12.4) | 81.9 (11.3) | 81.3 (10.5) | 80.6 (11.7) | 82.2 (10.5) | 90.9 (7.7) | 90.2 (7.4) | 90.1 (8.0) | 92.0 (7.9) | 97.8 (5.0) | 96.6 (5.4) | 95.4 (7.5) | 95.2 (5.8) |
| PPV of P (SD) | 62.6 (7.9) | 67.2 (7.5) | 40.9 (2.7) | 41.1 (2.2) | 41.3 (2.8) | 41.2 (2.3) | 39.1 (1.5) | 39.4 (1.5) | 39.5 (1.7) | 39.6 (1.4) | 38.5 (1) | 38.5 (1.3) | 38.9 (1.6) | 38.5 (1.2) |
| PPV of M (SD) | 46.2 (12.1) | 45.9 (9.9) | 24.8 (3.0) | 25.0 (2.7) | 25.1 (3.0) | 24.7 (2.5) | 22.7 (1.7) | 22.9 (1.5) | 22.9 (1.7) | 22.6 (1.7) | 21.5 (1.0) | 21.7 (1.0) | 21.9 (1.7) | 21.9 (1.2) |
| **Mice-Elasticnet** | | | | | | | | | | | | | | |
| SEN of P (SD) | 92.7 (3.3) | 93.7 (1.8) | 96.7 (3.4) | 95.9 (3.2) | 96.9 (3.5) | 97.5 (3.2) | 97.8 (3.2) | 97.1 (3.3) | 98.5 (2.9) | 99.3 (2) | | | | |
| SEN of M (SD) | 97.3 (7.7) | 99.8 (2.0) | 99.8 (2.0) | 100(0) | 99.8 (2.5) | 100 (0) | 99.8 (2.5) | 100 (0) | 99.9 (1.4) | 100 (0) | | | | |
| FPR of P (SD) | 40.9 (14.9) | 36.3 (11.5) | 92.8 (7.4) | 87.1 (7.5) | 92.5 (7.5) | 87.9 (7.4) | 97.5 (3.6) | 94.1 (5) | 97.5 (3.6) | 95.6 (4.7) | | | | |
| FPR of M (SD) | 38.8 (17.2) | 34.3 (12.7) | 90.2 (9.7) | 82.5 (10.6) | 89.9 (9.8) | 83.0 (10.5) | 96.0 (5.6) | 91.0 (7.5) | 96.1 (5.6) | 93.1 (7.4) | | | | |
| PPV of P (SD) | 59.9 (8.7) | 62.7 (7.5) | 39.5 (2.1) | 40.9 (2.3) | 39.6 (2.1) | 41 (2.3) | 38.5 (1.1) | 39.2 (1.5) | 38.7 (1.1) | 39.4 (1.3) | | | | |
| PPV of M (SD) | 43.0 (12.0) | 45.6 (9.9) | 23.0 (2.2) | 24.7 (2.7) | 23.0 (2.3) | 24.6 (2.5) | 21.7 (1.1) | 22.8 (1.5) | 21.8 (1.1) | 22.4 (1.5) | | | | |
| **MissForest-Lasso** | | | | | | | | | | | | | | |
| SEN of P (SD) | 92.3 (3.3) | 100 (1.8) | 90.2 (4.4) | 93.8 (2.3) | 91 (4.5) | 93.9 (2.4) | 88.6 (5.3) | 93.8 (2.6) | 90.2 (5) | 94.2 (2.5) | | | | |
| SEN of M (SD) | 96.6 (8.6) | 99.8 (2.0) | 92.8 (11.5) | 99.3 (4.3) | 93.2 (11.2) | 99.2 (4.5) | 90.0 (13.3) | 98.8 (5.5) | 94.5 (10.8) | 99.4 (3.8) | | | | |
| FPR of P (SD) | 35.9 (11.9) | 51.3 (11.1) | 36.7 (10.8) | 38.3 (10.9) | 37.6 (12.4) | 40.9 (11.7) | 36 (11.1) | 40.2 (11.2) | 40.1 (12.1) | 46.7 (11.9) | | | | |
| FPR of M (SD) | 33.4 (14.4) | 33.9 (12.4) | 32.7 (12.8) | 34.6 (12.3) | 33.9 (14.1) | 36.8 (12.3) | 31.9 (12.4) | 35.8 (12.9) | 35.4 (13.6) | 41.8 (13.6) | | | | |
| PPV of P (SD) | 62.6 (7.9) | 67.2 (7.5) | 62.8 (7.3) | 62.7 (6.7) | 62.4 (7.8) | 61.4 (6.9) | 62.9 (7.2) | 61.8 (7) | 61.7 (7.8) | 59.3 (6.7) | | | | |
| PPV of M (SD) | 46.2 (12.1) | 45.9 (9.9) | 45.1 (10.7) | 45.4 (10.4) | 44.6 (11.4) | 43.5 (9.1) | 44.9 (11.0) | 44.2 (9.7) | 43.9 (12.2) | 40.4 (8.6) | | | | |
| **100-covariate** (N=500) | $\rho = 0.2$ | $\rho = 0.8$ | | | | | $\rho = 0.2$ | $\rho = 0.8$ | $\rho = 0.2$ | $\rho = 0.8$ | | | | |
| **Mice-Lasso** | | | | | | | | | | | | | | |
| SEN of P (SD) | 88.9 (4.8) | 71.9 (6.2) | | | | | 94.7 (3) | 93.5 (4.9) | 94 (2.4) | 93.5 (4.3) | | | | |
| SEN of M (SD) | 88.5 (12.6) | 59.5 (15) | | | | | 99.5 (3) | 93.9 (10.4) | 99.6 (2.8) | 95.3 (9.1) | | | | |
| FPR of P (SD) | 18.6 (4.5) | 11.6 (3.6) | | | | | 100.2 (0.3) | 97.5 (1.1) | 100.1 (0.4) | 97.3 (1.3) | | | | |
| FPR of M (SD) | 14.6 (4.8) | 7.4 (3.6) | | | | | 99.5 (0.6) | 94.3 (2.2) | 99.4 (0.7) | 94.1 (2.4) | | | | |
| PPV of P (SD) | 24.7 (4.5) | 30.5 (6.2) | | | | | 5.7 (0.2) | 5.8 (0.3) | 5.7 (0.1) | 5.8 (0.2) | | | | |
| PPV of M (SD) | 21.6 (6.1) | 28 (10.8) | | | | | 4 (0.1) | 4 (0.4) | 4 (0.1) | 4.1 (0.4) | | | | |
| **Mice-Elasticnet** | | | | | | | | | | | | | | |
| SEN of P (SD) | 89 (4.8) | 72.3 (5.9) | | | | | 99.2 (2.2) | 99.6 (1.5) | 99 (2.3) | 99.2 (2.2) | | | | |
| SEN of M (SD) | 88.8 (12.6) | 60.3 (15) | | | | | 100 (0) | 99.9 (1.2) | 100 (0) | 100 (0) | | | | |
| FPR of P (SD) | 18.8 (4.6) | 12.1 (3.8) | | | | | 100 (0) | 100.0 (0.1) | 100 (0) | 100.0 (0.3) | | | | |
| FPR of M (SD) | 14.9 (5) | 7.8 (3.8) | | | | | 100 (0) | 100 (0.2) | 100 (0) | 99.9 (0.5) | | | | |
| PPV of P (SD) | 24.5 (4.7) | 29.7 (6.4) | | | | | 6 (0.1) | 6 (0.1) | 6 (0.1) | 6 (0.1) | | | | |
| PPV of M (SD) | 21.5 (6) | 27.4 (11.1) | | | | | 4 (0) | 4 (0) | 4 (0) | 4 (0) | | | | |
| **MissForest-Lasso** | | | | | | | | | | | | | | |
| SEN of P (SD) | 88.9 (4.8) | 71.9 (6.2) | | | | | 65.9 (6.5) | 64.7 (7) | 63.9 (6.6) | 64.8 (6.6) | | | | |
| SEN of M (SD) | 88.5 (12.6) | 59.5 (15) | | | | | 46.3 (11.5) | 45.4 (14.7) | 42.5 (10) | 42.5 (15) | | | | |
| FPR of P (SD) | 18.6 (4.5) | 11.6 (3.6) | | | | | 11.1 (4) | 11.7 (3.1) | 11.7 (4.1) | 11.5 (3) | | | | |
| FPR of M (SD) | 14.6 (4.8) | 7.4 (3.6) | | | | | 8 (3.8) | 7 (3.1) | 7.8 (3.4) | 7.1 (3) | | | | |
| PPV of P (SD) | 24.7 (4.5) | 30.5 (6.2) | | | | | 30.2 (7.5) | 27.8 (5.4) | 28.4 (7.1) | 28.3 (5.5) | | | | |
| PPV of M (SD) | 21.6 (6.1) | 28 (10.8) | | | | | 22.5 (11.5) | 23.6 (10.9) | 21.5 (11.2) | 22 (10.3) | | | | |
| **MissForest-Elasticnet** | | | | | | | | | | | | | | |
| SEN of P (SD) | 89 (4.8) | 72.3 (5.9) | | | | | 66 (6.9) | 65.9 (7.4) | 64.4 (6.6) | 65.1 (6.8) | | | | |
| SEN of M (SD) | 88.8 (12.6) | 60.3 (15) | | | | | 46.5 (11.4) | 47.7 (14.7) | 43.1 (10.4) | 43.5 (14.9) | | | | |
| FPR of P (SD) | 18.8 (4.6) | 12.1 (3.8) | | | | | 11.8 (4.1) | 12.4 (3.8) | 12.2 (4.1) | 11.9 (3.2) | | | | |
| FPR of M (SD) | 14.9 (5) | 7.8 (3.8) | | | | | 8.6 (4) | 7.4 (3.1) | 8.3 (3.7) | 7.3 (3) | | | | |
| PPV of P (SD) | 24.5 (4.7) | 29.7 (6.4) | | | | | 28.7 (6.7) | 27.3 (5.8) | 27.6 (6.7) | 27.6 (5.8) | | | | |
| PPV of M (SD) | 21.5 (6) | 27.4 (11.1) | | | | | 21 (9) | 23.3 (9.7) | 20.5 (10) | 22.2 (11.2) | | | | |

Table A.18: Comparison of average sensitivity (SEN), false positive rate (FPR) and positive predictive value (PPV) of selection for the predictors (P) and for the moderators (M) for the 1 SE tolerance models in the simulation study. Average SEN, FPR and PPV are given in percentages with corresponding SD.

| | **Selection of predictors and moderators in the 1SE tolerance models** | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Scenario 3** | | **Scenario 4: complete outcome** | | | | **Scenario 5: incomplete outcome (20%)** | | | | **Scenario 6: interactions in imputation model** | | | |
| Data | Complete | | MCAR | | MAR | | MCAR | | MAR | | MCAR | | MAR | |
| **20-covariate** | N=250 | N=1000 | N=250 | N=1000 | N=250 | N=1000 | N=250 | N=1000 | N=250 | N=1000 | N=250 | N=1000 | N=250 | N=1000 |
| **Mice-Lasso** | | | | | | | | | | | | | | |
| SEN of P (SD) | 91.7 (3.5) | 100 (1.3) | 94.4 (3.2) | 94.9 (2.9) | 94.8 (3.5) | 96.5 (3.3) | 95.1 (3.5) | 96 (3.3) | 96.6 (3.7) | 99 (2.4) | 97.2 (3.5) | 95.5 (3.1) | 96.4 (3.6) | 95.5 (3.1) |
| SEN of M (SD) | 96.3 (8.9) | 99.8 (2.0) | 99.1 (4.7) | 99.9 (1.4) | 99.2 (4.5) | 100 (0) | 99.5 (3.5) | 100 (0) | 99.3 (4.0) | 100 (0) | 99.8 (2.5) | 99.9 (1.44) | 99.4 (3.8) | 100 (0) |
| FPR of P (SD) | 25.6 (10.2) | 25.2 (9.1) | 72.3 (12.1) | 66.7 (10) | 71.4 (12.2) | 68.5 (10) | 84.4 (8.5) | 79.6 (7.4) | 84.3 (8.8) | 82.5 (7.6) | 93.5 (9.4) | 85.4 (10.3) | 88.7 (11.6) | 83.3 (9.6) |
| FPR of M (SD) | 24.2 (12.1) | 22.1 (10.4) | 65.3 (13.7) | 60.2 (12.4) | 64.4 (14.7) | 62.3 (12.7) | 77.5 (11.7) | 71.6 (10.5) | 77.2 (12.3) | 75.2 (10.3) | 92.1 (10.9) | 82.6 (12.2) | 86.1 (13.7) | 79.9 (11.6) |
| PPV of P (SD) | 70.2 (8.6) | 81 (8.2) | 45.3 (4.5) | 47.4 (3.9) | 45.7 (4.5) | 47.1 (3.8) | 41.5 (2.7) | 43.1 (2.5) | 41.9 (2.8) | 43 (2.4) | 39.5 (2.6) | 41.3 (3.1) | 40.7 (3.5) | 41.9 (2.9) |
| PPV of M (SD) | 54.4 (13.6) | 57.2 (12.8) | 29.5 (4.9) | 31.4 (4.9) | 30.0 (5.6) | 30.6 (4.7) | 25.8 (3.4) | 27.5 (3.1) | 25.0 (3.6) | 26.5 (2.9) | 22.6 (2.6) | 24.7 (3.0) | 24.0 (3.5) | 25.4 (3.0) |
| **Mice-Elasticnet** | | | | | | | | | | | | | | |
| SEN of P (SD) | 89.3 (5.1) | 93.3 (1) | 94.7 (4.1) | 93.6 (1.4) | 95.4 (4.1) | 94 (2.4) | 95.8 (3.7) | 94 (2.2) | 97.2 (3.9) | 97 (3.3) | | | | |
| SEN of M (SD) | 95.9 (9.5) | 99.8 (2.0) | 99.3 (4.3) | 100(0) | 99.4 (3.8) | 100 (0) | 99.4 (3.8) | 100 (0) | 99.7 (2.9) | 100 (0) | | | | |
| FPR of P (SD) | 18.2 (16.7) | 8.4 (4.9) | 74.7 (20.1) | 30.8 (13) | 74.5 (20.7) | 32.7 (14.6) | 86 (13.6) | 43 (15.9) | 87 (14.3) | 51.1 (18.7) | | | | |
| FPR of M (SD) | 19.3 (17.0) | 10.8 (6.6) | 71.9 (20.8) | 29.0 (13.7) | 71.4 (21.5) | 32.8 (14.8) | 82.0 (16.4) | 39.4 (15.8) | 83.1 (17.4) | 48.4 (18.1) | | | | |
| PPV of P (SD) | 78.6 (13.6) | 87.9 (6.3) | 45.4 (7.5) | 66.7 (8.1) | 45.7 (8.1) | 65.6 (8.4) | 41.4 (4.2) | 59 (8.3) | 41.6 (4.6) | 55.7 (8.7) | | | | |
| PPV of M (SD) | 63.5 (18.8) | 73.4 (13.2) | 28.4 (7.4) | 50.4 (10.8) | 28.8 (8.4) | 47.2 (9.8) | 25.1 (4.6) | 42.6 (9.8) | 25.0 (5.2) | 37.6 (8.8) | | | | |
| **MissForest-Lasso** | | | | | | | | | | | | | | |
| SEN of P (SD) | 91.7 (3.5) | 100 (1.3) | 89.1 (4.8) | 93.5 (1.8) | 90.2 (4.6) | 93.6 (2.1) | 86 (6.3) | 93.4 (2.1) | 88.4 (5.9) | 93.8 (2.3) | | | | |
| SEN of M (SD) | 96.6 (8.6) | 99.8 (2.0) | 91.9 (12.1) | 99.4 (3.8) | 92.7 (11.4) | 99.3 (4.3) | 88.0 (14.1) | 99.0 (4.9) | 93.7 (11.63) | 99.6 (3.2) | | | | |
| FPR of P (SD) | 25.6 (10.2) | 25.2 (9.1) | 26.3 (9.1) | 24.2 (9.4) | 27.6 (10.9) | 27.2 (9.6) | 25.8 (10.5) | 26.1 (9.5) | 29.9 (10.6) | 33.3 (10.2) | | | | |
| FPR of M (SD) | 24.2 (12.1) | 22.1 (10.4) | 23.5 (10.3) | 22.4 (10.7) | 25.0 (12.0) | 25.0 (10.9) | 23.1 (11.1) | 23.8 (10.7) | 26.4 (11.5) | 30.3 (11.2) | | | | |
| PPV of P (SD) | 70.2 (8.6) | 81 (8.2) | 68.8 (7.8) | 71.6 (8.1) | 68.3 (8.7) | 69.1 (7.7) | 68.8 (8.7) | 70 (8.1) | 65.8 (8.2) | 64.6 (7.4) | | | | |
| PPV of M (SD) | 54.4 (13.6) | 57.2 (12.8) | 53.21 (12.2) | 56.8 (12.9) | 52.5 (13.7) | 53.7 (11.3) | 52.94 (13.3) | 55.1 (12.3) | 50.9 (12.6) | 48.7 (10.1) | | | | |
| **100-covariate** (N=500) | $\rho = 0.2$ | $\rho = 0.8$ | | | | | $\rho = 0.2$ | $\rho = 0.8$ | $\rho = 0.2$ | $\rho = 0.8$ | | | | |
| **Mice-Lasso** | | | | | | | | | | | | | | |
| SEN of P (SD) | 86 (5.8) | 68.7 (6) | | | | | 94.6 (3) | 92.7 (5.1) | 93.8 (2.6) | 92.5 (4.4) | | | | |
| SEN of M (SD) | 82.8 (13.5) | 53.3 (13.5) | | | | | 99.4 (3.4) | 92.1 (11.4) | 99.1 (4.1) | 93.3 (10.1) | | | | |
| FPR of P (SD) | 13.2 (3.7) | 8.3 (2.8) | | | | | 100.0 (0.4) | 96.7 (1.3) | 100.0 (0.4) | 96.3 (1.6) | | | | |
| FPR of M (SD) | 9.9 (3.8) | 4.7 (2.7) | | | | | 99.3 (0.8) | 92.8 (2.5) | 99.2 (0.9) | 92.1 (2.9) | | | | |
| PPV of P (SD) | 31.3 (6.1) | 37.5 (8) | | | | | 5.7 (0.2) | 5.8 (0.3) | 5.7 (0.1) | 5.8 (0.3) | | | | |
| PPV of M (SD) | 28.1 (8.4) | 36.5 (15.4) | | | | | 4 (0.1) | 4 (0.5) | 4 (0.2) | 4.1 (0.4) | | | | |
| **Mice-Elasticnet** | | | | | | | | | | | | | | |
| SEN of P (SD) | 85.8 (5.8) | 67.2 (6.3) | | | | | 99 (2.4) | 99.1 (2.4) | 98.7 (2.7) | 98.5 (2.9) | | | | |
| SEN of M (SD) | 82.6 (13) | 52.2 (12.1) | | | | | 100 (0) | 99.8 (2) | 100 (0) | 99.7 (2.3) | | | | |
| FPR of P (SD) | 12.8 (3.9) | 7 (2.4) | | | | | 100 (0) | 100.0 (0.3) | 100 (0) | 100.0 (0.9) | | | | |
| FPR of M (SD) | 9.5 (3.9) | 3.8 (2.5) | | | | | 100 (0.1) | 99.7 (0.6) | 100 (0.1) | 99.4 (1.6) | | | | |
| PPV of P (SD) | 32.2 (7.1) | 41.3 (8.4) | | | | | 6 (0.1) | 6 (0.1) | 5.9 (0.2) | 6 (0.2) | | | | |
| PPV of M (SD) | 29.4 (9.9) | 42.4 (18.4) | | | | | 4 (0) | 4 (0.1) | 4 (0) | 4 (0.1) | | | | |
| **MissForest-Lasso** | | | | | | | | | | | | | | |
| SEN of P (SD) | 86 (5.8) | 68.7 (6) | | | | | 62 (6.5) | 62.4 (6.9) | 60.5 (6.3) | 62.1 (7.1) | | | | |
| SEN of M (SD) | 82.8 (13.5) | 53.3 (13.5) | | | | | 43.3 (9.7) | 43.7 (12.9) | 40.9 (8.6) | 40.8 (13.7) | | | | |
| FPR of P (SD) | 13.2 (3.7) | 8.3 (2.8) | | | | | 7 (2.9) | 8.6 (2.3) | 7.3 (3) | 8.5 (2.3) | | | | |
| FPR of M (SD) | 9.9 (3.8) | 4.7 (2.7) | | | | | 4.6 (2.9) | 4.9 (2.3) | 4.5 (2.6) | 5.2 (2.4) | | | | |
| PPV of P (SD) | 31.3 (6.1) | 37.5 (8) | | | | | 40.3 (10.1) | 34 (6.8) | 38.6 (10.1) | 34.2 (6.9) | | | | |
| PPV of M (SD) | 28.1 (8.4) | 36.5 (15.4) | | | | | 34.2 (17.5) | 30.5 (14.1) | 33.6 (18.3) | 27.6 (13.1) | | | | |
| **MissForest-Elasticnet** | | | | | | | | | | | | | | |
| SEN of P (SD) | 85.8 (5.8) | 67.2 (6.3) | | | | | 53.3 (7) | 57.1 (8.5) | 50.6 (7.5) | 57.6 (7.3) | | | | |
| SEN of M (SD) | 82.6 (13) | 52.2 (12.1) | | | | | 41.9 (9) | 48.3 (12.9) | 39.2 (9.8) | 44.3 (13.3) | | | | |
| FPR of P (SD) | 12.8 (3.9) | 7 (2.4) | | | | | 2.5 (1.5) | 5.7 (2.4) | 2.6 (1.7) | 5.7 (1.7) | | | | |
| FPR of M (SD) | 9.5 (3.9) | 3.8 (2.5) | | | | | 1.3 (1.4) | 2.8 (2.1) | 1.2 (1.4) | 3.1 (2.1) | | | | |
| PPV of P (SD) | 32.2 (7.1) | 41.3 (8.4) | | | | | 66.2 (15.1) | 43.3 (9.4) | 64.2 (15.4) | 42.7 (8.9) | | | | |
| PPV of M (SD) | 29.4 (9.9) | 42.4 (18.4) | | | | | 66.3 (24.5) | 49.8 (23.7) | 68.7 (27.4) | 44.2 (22.8) | | | | |

Table A.19: Comparison of average sensitivity (SEN), false positive rate (FPR) and positive predictive value (PPV) of selection for the predictors (P) and for the moderators (M) for the 15% tolerance models in the simulation study. Average SEN, FPR and PPV are given in percentages with corresponding SD.

| | **Selection of predictors and moderators in the 15% tolerance models** | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Scenario 3** | | **Scenario 4: complete outcome** | | | | **Scenario 5: incomplete outcome (20%)** | | | | **Scenario 6: interactions in imputation model** | | | |
| Data | Complete | | MCAR | | MAR | | MCAR | | MAR | | MCAR | | MAR | |
| **20-covariate** | N=250 | N=1000 | N=250 | N=1000 | N=250 | N=1000 | N=250 | N=1000 | N=250 | N=1000 | N=250 | N=1000 | N=250 | N=1000 |
| **Mice-Lasso** | | | | | | | | | | | | | | |
| SEN of P (SD) | 84.2 (6) | 91.9 (3.2) | 86.3 (6.9) | 89.3 (3.3) | 86.8 (5.6) | 89.4 (3.3) | 87.5 (6.4) | 89.6 (3.6) | 88.9 (6.1) | 90.4 (3.5) | 88.7 (5.8) | 89.9 (3.4) | 88.1 (6.5) | 89.5 (3.4) |
| SEN of M (SD) | 91.8 (13.0) | 99.5 (3.5) | 94.2 (11.7) | 99.8 (2.0) | 95.2 (10.3) | 100 (0) | 93.5 (11.4) | 99.3 (4.3) | 95.3 (10.0) | 99.92 (1.4) | 96.4 (9.2) | 99.8 (2.5) | 96.1 (10.0) | 99.8 (2.5) |
| FPR of P (SD) | 5.9 (4.6) | 0.0 (3.2) | 20.2 (9.2) | 10.8 (5.2) | 19.7 (8.8) | 12 (5.2) | 30.2 (11.5) | 13.9 (5.4) | 31 (11.6) | 17.6 (5.2) | 30.9 (17.2) | 10.7 (5.6) | 24.5 (10.9) | 10.9 (5.5) |
| FPR of M (SD) | 7.0 (6.1) | 4.0 (4.9) | 18.2 (10.2) | 10.6 (7.1) | 18.4 (9.8) | 12.2 (7.2) | 26.4 (12.2) | 13.7 (7.5) | 27.9 (11.6) | 19.7 (7.6) | 30.3 (17.1) | 11.1 (7.3) | 24.5 (11.5) | 11.7 (7.0) |
| PPV of P (SD) | 90.4 (6.7) | 100 (5) | 74 (8.9) | 84.3 (6.5) | 74.4 (8.2) | 82.8 (6.3) | 65.6 (8.2) | 80.6 (6.2) | 65.4 (8.3) | 76.7 (5.2) | 66.5 (11.1) | 84.6 (7.1) | 70.4 (8.9) | 84.3 (6.9) |
| PPV of M (SD) | 80.1 (15) | 88.7 (12.7) | 61.2 (14.9) | 74.2 (14) | 60.9 (13.7) | 71.1 (13.5) | 51.5 (13.3) | 68.3 (13.3) | 49.9 (11.5) | 59.1 (9.9) | 49.8 (14) | 73.2 (14.3) | 53.7 (12.5) | 71.9 (13.6) |
| **Mice-Elasticnet** | | | | | | | | | | | | | | |
| SEN of P (SD) | 41.9 (27.4) | 20.3 (5.2) | 84.3 (13.6) | 32.8 (21.1) | 84.8 (15.9) | 36.7 (23.2) | 87 (11.4) | 42.5 (26.1) | 89.6 (11.9) | 57.4 (27.3) | | | | |
| SEN of M (SD) | 48.2 (34.6) | 24.5 (13.2) | 92.3 (15.0) | 42.3 (24.8) | 92.5 (15.8) | 47.5 (26.6) | 93.0 (15.2) | 47.9 (31.4) | 94.7 (14.8) | 66.2 (33.2) | | | | |
| FPR of P (SD) | 5.1 (8.9) | 0.1 (0.7) | 34 (18.9) | 2.4 (5.4) | 35 (19.3) | 3.2 (6.5) | 43.2 (20) | 4.4 (7.2) | 47.6 (21) | 9.4 (10.3) | | | | |
| PPV of M (SD) | 6.3 (9.7) | 0.2 (1.1) | 34.0 (18.4) | 3.1 (6.5) | 35.6 (18.8) | 4.2 (7.8) | 41.3 (18.9) | 5.4 (7.7) | 45.4 (19.6) | 11.5 (11.2) | | | | |
| PPV of P (SD) | 90.6 (11.5) | 99.3 (4) | 64.4 (13) | 94.9 (8.8) | 64 (13.1) | 93.6 (9.5) | 58.7 (12.1) | 91.6 (10.2) | 57 (11.8) | 85.1 (11.5) | | | | |
| PPV of M (SD) | 78.1 (21.6) | 98.2 (10) | 46.6 (15.1) | 89.2 (17.4) | 45.7 (14.7) | 86.9 (18.1) | 41 (12.7) | 80.8 (19.6) | 38.6 (12.3) | 68.5 (19.2) | | | | |
| **MissForest-Lasso** | | | | | | | | | | | | | | |
| SEN of P (SD) | 84.2 (6) | 91.9 (3.2) | 71.8 (10.8) | 86.1 (3.6) | 73.5 (9.9) | 86.4 (3.1) | 62 (12.3) | 83.8 (4.6) | 67.9 (11.8) | 85.6 (3.1) | | | | |
| SEN of M (SD) | 91.8 (13.0) | 99.5 (3.5) | 77.7 (18.0) | 96.4 (9.5) | 80.2 (18.1) | 97.4 (7.6) | 66.1 (20.9) | 93.3 (12.5) | 79.2 (15.2) | 98.3 (6.7) | | | | |
| FPR of P (SD) | 5.9 (4.6) | 0 (3.2) | 5.4 (4.3) | 2.7 (3) | 6 (4.6) | 3.5 (3.7) | 4.8 (4.4) | 3.4 (3.3) | 6.7 (5) | 5.3 (4) | | | | |
| FPR of M (SD) | 7.0 (6.1) | 4.0 (4.9) | 5.8 (5.4) | 3.5 (4.3) | 6.0 (5.8) | 4.1 (4.9) | 5 (5.4) | 3.9 (4.5) | 7.1 (6.3) | 6.5 (5.3) | | | | |
| PPV of P (SD) | 90.4 (6.7) | 100 (5) | 89.9 (7.2) | 95.5 (4.9) | 89.1 (7.6) | 94.3 (5.7) | 89.9 (8.4) | 94.3 (5.4) | 87.2 (8.4) | 91.3 (6.1) | | | | |
| PPV of M (SD) | 80.1 (15) | 88.7 (12.7) | 81.1 (16.3) | 89.7 (12.1) | 81 (17) | 88.2 (13) | 81.2 (18.8) | 88.4 (12.7) | 77.6 (17.6) | 82.1 (13) | | | | |
| **100-covariate** (N=500) | $\rho = 0.2$ | $\rho = 0.8$ | | | | | $\rho = 0.2$ | $\rho = 0.8$ | $\rho = 0.2$ | $\rho = 0.8$ | | | | |
| **Mice-Lasso** | | | | | | | | | | | | | | |
| SEN of P (SD) | 63.2 (8.2) | 51 (6.7) | | | | | 83.7 (6.5) | 77.1 (5.2) | 83.6 (7) | 76.8 (6.3) | | | | |
| SEN of M (SD) | 52.6 (14.1) | 55.9 (8.6) | | | | | 73.3 (17.8) | 58.1 (11.2) | 74.3 (17.3) | 58.1 (12.8) | | | | |
| FPR of P (SD) | 0.8 (0.5) | 1.2 (0.7) | | | | | 83.2 (6.8) | 47.1 (9.1) | 82.7 (6.7) | 39.3 (9.2) | | | | |
| FPR of M (SD) | 0.2 (0.4) | 0.1 (0.3) | | | | | 70.5 (9.7) | 30.1 (9) | 69.7 (9.4) | 24.5 (8.5) | | | | |
| PPV of P (SD) | 92.8 (8.8) | 82.9 (13.4) | | | | | 6.1 (0.6) | 9.9 (1.8) | 6.1 (0.6) | 11.7 (2.4) | | | | |
| PPV of M (SD) | 94 (12.6) | 96 (10.6) | | | | | 4.2 (1.1) | 8.2 (3.1) | 4.3 (1) | 10 (4) | | | | |
| **Mice-Elasticnet** | | | | | | | | | | | | | | |
| SEN of P (SD) | 36 (8) | 31.3 (6.4) | | | | | 92.8 (4.7) | 87.2 (5.5) | 92.1 (4.8) | 86.9 (5.8) | | | | |
| SEN of M (SD) | 37 (10.2) | 28.8 (11) | | | | | 95.4 (9.5) | 81.4 (14.1) | 94.3 (10.8) | 81.7 (14.2) | | | | |
| FPR of P (SD) | 0.4 (0) | 0.5 (0.2) | | | | | 96.6 (5.4) | 79.9 (10.9) | 95.9 (5.6) | 73.2 (12.6) | | | | |
| FPR of M (SD) | 0 (0) | 0 (0.1) | | | | | 93.4 (8.2) | 68.5 (14.5) | 92.4 (8.3) | 60.8 (15.6) | | | | |
| PPV of P (SD) | 100 (0.7) | 97.9 (6) | | | | | 5.8 (0.4) | 6.7 (0.9) | 5.8 (0.4) | 7.3 (1.3) | | | | |
| PPV of M (SD) | NA | NA | | | | | 4.1 (0.4) | 5 (1.4) | 4.1 (0.5) | 5.7 (2.1) | | | | |
| **MissForest-Lasso** | | | | | | | | | | | | | | |
| SEN of P (SD) | 63.2 (8.2) | 51 (6.7) | | | | | 36.2 (7) | 38.1 (7.3) | 31.8 (8) | 37.4 (7.7) | | | | |
| SEN of M (SD) | 52.6 (14.1) | 55.9 (8.6) | | | | | 28.6 (15.1) | 44.3 (14.2) | 22.5 (15) | 40.2 (14.1) | | | | |
| FPR of P (SD) | 0.8 (0.5) | 1.2 (0.7) | | | | | 0.6 (0.3) | 1.9 (1) | 0.6 (0.3) | 1.8 (0.9) | | | | |
| FPR of M (SD) | 0.2 (0.4) | 0.1 (0.3) | | | | | 0.1 (0.3) | 0.4 (0.7) | 0.1 (0.3) | 0.5 (0.8) | | | | |
| PPV of P (SD) | 92.8 (8.8) | 82.9 (13.4) | | | | | 94.5 (9) | 66.6 (16.7) | 93.6 (10.5) | 67.5 (16.7) | | | | |
| PPV of M (SD) | 94 (12.6) | 96 (10.6) | | | | | 93.3 (17) | 87.7 (19.1) | 93.4 (19.3) | 85.5 (21.6) | | | | |
| **MissForest-Elasticnet** | | | | | | | | | | | | | | |
| SEN of P (SD) | 36 (8) | 31.3 (6.4) | | | | | 0.1 (0.8) | 13.4 (12.3) | 0.2 (2) | 10.8 (9.7) | | | | |
| SEN of M (SD) | 37 (10.2) | 28.8 (11) | | | | | 0 (0) | 8.7 (15.2) | 0.3 (2.8) | 5.7 (11.6) | | | | |
| FPR of P (SD) | 0.4 (0) | 0.5 (0.2) | | | | | 0.4 (0) | 1.6 (2.9) | 0.4 (0) | 0.8 (0.8) | | | | |
| FPR of M (SD) | 0 (0) | 0 (0.1) | | | | | 0 (0) | 0.2 (0.9) | 0 (0) | 0 (0.1) | | | | |
| PPV of P (SD) | 100 (0.7) | 97.9 (6) | | | | | NA | NA | NA | NA | | | | |
| PPV of M (SD) | NA | NA | | | | | NA | 89.4 (22.5) | 88.9 (19.2) | 98.5 (8.7) | | | | |

# Appendix B

# R code

## B.1 Musoro et al 2014 code error

Loading the caret package:

```
library(caret)
```

Loading the dataset (simulated dataset with 250 observations and 21 variables: 10 noise and 10 true predictors, and the outcome variable simulated as for Musoro et al. 2014)

```
Data<- read.csv("Simulations/CompleteData/normalSim.250_20.complete.1.csv", sep=";",
 dec=",",stringsAsFactors= TRUE)[,-1]
```

Fitting the Lasso with bootstrap tuning:

```
lassoGrid<-expand.grid(alpha = 1 , lambda= 10^seq(0.1,-1.8,length=40))
modelMatrix <- model.matrix(as.formula(paste(colnames(Data)[dim(Data)[2]],"~.",
                                    sep="")),data.frame(Data))
Outcome <- Data[,dim(Data)[2]]
set.seed(2)
(Fit.Caret <- train(modelMatrix[,-1],Outcome,method="glmnet",tuneGrid=lassoGrid,
                                    family="gaussian",
trControl=trainControl(number=100, method="boot", selectionFunction="best")))
```

So the best tuning parameter is saved in 'bestTune$lambda', in 'finalModel$lambdaOpt' and in 'results$lambda':

```
Fit.Caret$bestTune$lambda
## [1] 0.05443741
Fit.Caret$finalModel$lambdaOpt
## [1] 0.05443741
Fit.Caret$results$lambda[best(Fit.Caret$results,"RMSE",maximize=FALSE)]
## [1] 0.05443741
```

### B.1.1 Wrong commands for best and tolerance model coefficients (Musoro et al 2014)

1. The first wrong code line is the one computing the coefficients for the best $\lambda$ model:

```
coef(Fit.Caret$finalModel)[,best(Fit.Caret$results[nrow(Fit.Caret$results):1,],
        "RMSE",maximize=FALSE)]
## (Intercept)          V1          V2          V3          V4          V5
##   1.8243041   0.0000000   0.0000000   0.0000000   0.0000000   0.0000000
##          V6          V7          V8          V9         V10         V11
##  -0.7747893   0.8138004  -1.5656112   0.7144349   0.6548666   0.0000000
##         V12         V13         V14         V15         V16         V17
```

```
##    0.0000000    0.1453461    0.0000000    0.0000000    0.2892636   -0.5617533
##          V18          V19          V20
##   -0.5704481    1.5068930    0.7923402
```

The 'best' function has the correct argument, i.e. 'Fit.Caret$results', and it selects the row of the best $\lambda$ in the matrix 'Fit.Caret$results'. However this command is wrong because the row of the best $\lambda$ in the matrix 'Fit.Caret$results' is then extract as column from 'Fit.Caret$finalModel' that is a completely different object, for example, 'Fit.Caret$results' has 40 rows and 'Fit.Caret$finalModel' has 69 rows.

2. Similarly, the following command is wrong for the tolerance model (with MSE within 3% of the minimum) coefficients:

```
coef(Fit.Caret$finalModel)[,tolerance(Fit.Caret$results[nrow(Fit.Caret$results):
    1,], metric="RMSE",maximize=FALSE, tol=3)]
## (Intercept)          V1          V2          V3          V4          V5
##  2.42927035  0.00000000  0.00000000  0.00000000  0.00000000  0.00000000
##          V6          V7          V8          V9         V10         V11
## -0.30714608  0.37892981 -1.10390962  0.08959322  0.00000000  0.00000000
##         V12         V13         V14         V15         V16         V17
##  0.00000000  0.00000000  0.00000000  0.00000000  0.11480705 -0.36930415
##         V18         V19         V20
## -0.30683700  1.37039583  0.61817961
```

Compare the best $\lambda$ and the $\lambda$ to which the coefficients of the first wrong command (1) correspond to:

```
Fit.Caret$bestTune$lambda
## [1] 0.05443741
Fit.Caret$finalModel$lambda[best(Fit.Caret$results[nrow(Fit.Caret$results):1,],"RMSE",
    maximize=FALSE)]
## [1] 0.1665014
```

### B.1.2 Correct commands for best and tolerance model coefficients

1. The following command correctly selects the best $\lambda$ model coefficients:

```
coef(Fit.Caret$finalModel,s=Fit.Caret$bestTune$lambda)
## 21 x 1 sparse Matrix of class "dgCMatrix"
##                      1
## (Intercept)  1.51417057
## V1           0.10929752
## V2           0.10691587
## V3           .
## V4           0.04622961
## V5          -0.04722650
## V6          -0.91077606
## V7           0.99546539
## V8          -1.87870078
## V9           0.94710545
## V10          0.96779451
## V11         -0.00972630
## V12         -0.07203485
## V13          0.24297913
## V14         -0.02727955
## V15          0.07863696
## V16          0.36672550
## V17         -0.66139761
## V18         -0.68156684
## V19          1.54822298
## V20          0.88933726
```

2. For the 3% tolerance model coefficients, we first find the $\lambda$ closest to the 3% tolerance $\lambda$ in 'Fit.Caret$results' as 'caret' does not return the exact $\lambda$ with the 'tolerance' function (see below). Therefore, we need to find $\lambda$ such that

```
(tol.tune<-Fit.Caret$results$lambda[(Fit.Caret$results$RMSE-
min(Fit.Caret$results$RMSE))*100/min(Fit.Caret$results$RMSE)>=
3][Fit.Caret$results$lambda[(Fit.Caret$results$RMSE-
min(Fit.Caret$results$RMSE))*100/min(Fit.Caret$results$RMSE)>=3]==
min(Fit.Caret$results$lambda[(Fit.Caret$results$RMSE-
min(Fit.Caret$results$RMSE))*100/min(Fit.Caret$results$RMSE)>=3])])
## [1] 0.1869799
```

Then we find the correspondent tolerance coefficients:

```
coef(Fit.Caret$finalModel)[,Fit.Caret$finalModel$lambda==
        Fit.Caret$finalModel$lambda[abs(Fit.Caret$finalModel$
lambda-tol.tune)==min(abs(Fit.Caret$finalModel$lambda-tol.tune))]]
## (Intercept)          V1          V2          V3          V4          V5
##   1.8649046   0.0000000   0.0000000   0.0000000   0.0000000   0.0000000
##          V6          V7          V8          V9         V10         V11
##  -0.7467256   0.7862884  -1.5351462   0.6746960   0.6098690   0.0000000
##         V12         V13         V14         V15         V16         V17
##   0.0000000   0.1305558   0.0000000   0.0000000   0.2784649  -0.5490214
##         V18         V19         V20
##  -0.5540582   1.4982586   0.7807537
```

Compare the closest to tolerance lambda (3%) 'tol.tune' and the $\lambda$ which the coefficients of the second wrong command (see Subsection B.1.1 above) correspond to:

```
tol.tune
## [1] 0.1869799
Fit.Caret$finalModel$lambda[tolerance(Fit.Caret$results[nrow(Fit.Caret$results):
      1,], "RMSE",maximize=FALSE,tol=3)]
## [1] 0.4221418
```

Also, notice that the 'tolerance' function does not select the most parsimonious model if applied to 'Fit.Caret$results':

```
Fit.Caret$results$lambda[tolerance(Fit.Caret$results,metric="RMSE",maximize=FALSE,
      tol=3)]
## [1] 0.01584893
```

The returned tolerance lambda is even smaller than the best lambda, thus the correspondent model will not be more parsimonious than the best model.

## B.2 MissForest-Lasso R function

```
# FUNCTIONS NEEDED:_____________________________________________________

### 1st FUNCTION:_____________________________________________________
### A function to compute apparent performance ###################################
MSE.alpha.beta <- function (coef,data.imputed,Num.imputed) {
Alpha.apparent <- rep(NA,Num.imputed); Beta.apparent <- rep(NA,Num.imputed)
MSE.apparent   <- rep(NA,Num.imputed)

i=1
while(i <= Num.imputed){
Outcome.data <- data.imputed[[i]][,dim(data.imputed[[i]])[2]]
modelMatrix.data <- model.Matrix(as.formula(paste(colnames(data.imputed[[i]])[
dim(data.imputed[[i]])[2]],"~.%in%",colnames(data.imputed[[
i]])[1],"+.",sep="")),data.imputed[[i]])
X.data<-modelMatrix.data[,-1]
Predicted.data <- as.matrix(coef[1] + X.data%*%coef[-1])
MSE.apparent[i]    <- mean((Outcome.data-Predicted.data)^2)
lm.coef        <- lm(Outcome.data ~ Predicted.data)$coef
Alpha.apparent[i]  <- lm.coef[1]
Beta.apparent[i]       <- lm.coef[2]
```

```
i<-i+1
}
out.opti.best.tol  <- list(MSE.apparent,Alpha.apparent,Beta.apparent)
out.opti.best.tol
}


### 2nd FUNCTION:_____________________________________________________________
### MissForest-Lasso ##########################################################

MissForest<- function (Num.imputed,Num.boot=NULL,method,Num.cv=NULL,repeats=NULL,
Data.NA,Grid,percent.tol,cores_2_use,parallelize,seed){

###- Peforming missForest

data.imputed <- vector("list",Num.imputed)
Imput<- vector("list",Num.imputed)

j=1
while (j <= Num.imputed){

cl <- makeCluster(cores_2_use)
clusterExport(cl, c("Data.NA","parallelize","cores_2_use","seed","Imput"),
envir=environment())
clusterSetRNGStream(cl, seed)
clusterEvalQ(cl, library(missForest))
Imput[[j]] <-
parLapply(cl = cl, X = 1:cores_2_use, fun = function(no){
missForest(Data.NA, maxiter = 10, ntree = 100, variablewise = TRUE,
verbose = TRUE,
replace = TRUE,
parallelize = parallelize)
})
stopCluster(cl)
data.imputed[[j]]=Imput[[j]][[1]]$ximp
j<-j+1
}


###- Fit glmnet using caret.

modelMatrixData<-model.Matrix(as.formula(paste(colnames(Data.NA)[dim(Data.NA)[2
]],"~.%in%", colnames(Data.NA)[1],"+.",sep="")),Data.NA)
coef.best<-coef.tol.1SE<-coef.tol<-coef.tol.15<- matrix(NA,Num.imputed,
dim(modelMatrixData)[2])
lambda<- matrix(NA,Num.imputed,4)

i=1
while (i <= Num.imputed){

modelMatrix<-model.Matrix(as.formula(paste(colnames(data.imputed[[i]])[
dim(data.imputed[[i]])[2]],"~.%in%", colnames(data.imputed[[i
]])[1],"+.",sep="")),data.imputed[[i]])
Outcome    <- data.imputed[[i]][,dim(data.imputed[[i]])[2]]

options(warn=-1)

cl=makeCluster(cores_2_use);registerDoParallel(cl)
if (method=="repeatedcv")
{set.seed(seed)
```

```
Fit.Caret <- train(modelMatrix[,-1] ,Outcome, method="glmnet",tuneGrid=Grid,
family="gaussian", trControl=trainControl(number=Num.cv,
method=method, repeats=repeats, selectionFunction="best"))
} else if (method=="boot")
{set.seed(seed)
Fit.Caret <- train(modelMatrix[,-1] ,Outcome, method="glmnet",tuneGrid=Grid,
family="gaussian",trControl=trainControl(number=Num.boot,
method=method, selectionFunction="best"))
}
stopCluster(cl)
options(warn=0)

#- Tuning parameters

lambda[i,1] <- Fit.Caret$finalModel$lambdaOpt
lambda[i,2] <- max(Fit.Caret$results$lambda[Fit.Caret$results$RMSE<=
Fit.Caret$results[row.names(Fit.Caret$bestTune),]$RMSE +
(Fit.Caret$results[row.names(Fit.Caret$bestTune),]$RMSESD)/
sqrt(Num.boot)])
lambda[i,3] <- max(Fit.Caret$results$lambda[(Fit.Caret$results$RMSE-
min(Fit.Caret$results$RMSE))*100/min(Fit.Caret$results$RMSE)<=
percent.tol])
lambda[i,4] <- max(Fit.Caret$results$lambda[(Fit.Caret$results$RMSE-
min(Fit.Caret$results$RMSE))*100/min(Fit.Caret$results$RMSE
)<=15])

#- Model coefficients -#
coef.best[i,]    <- as.matrix(coef(Fit.Caret$finalModel,s=
Fit.Caret$bestTune$lambda))
coef.tol.1SE[i,]<- as.matrix(coef(Fit.Caret$finalModel,s=
max(Fit.Caret$results$lambda[Fit.Caret$results$RMSE<=
Fit.Caret$results[row.names(Fit.Caret$bestTune),]$RMSE +
(Fit.Caret$results[row.names(Fit.Caret$bestTune),]$RMSESD)/
sqrt(Num.boot)])))
coef.tol[i,]     <-as.matrix(coef(Fit.Caret$finalModel,s=max(Fit.Caret$results$
lambda[(Fit.Caret$results$RMSE-min(Fit.Caret$results$RMSE))*
100/min(Fit.Caret$results$RMSE)<=percent.tol])))
coef.tol.15[i,] <-as.matrix(coef(Fit.Caret$finalModel,s=max(Fit.Caret$results$
lambda[(Fit.Caret$results$RMSE-min(Fit.Caret$results$RMSE))*
100/min(Fit.Caret$results$RMSE)<=15])))
i<-i+1
}

Av.coef.best                <-apply(coef.best,2,mean )
Av.coef.tol.1SE             <-apply(coef.tol.1SE,2,mean )
Av.coef.tol                 <-apply(coef.tol,2,mean)
Av.coef.tol.15              <-apply(coef.tol.15,2,mean)
if(Num.imputed==1){
Av.lambda                  <- lambda
} else {
Av.lambda                  <- apply(lambda,2,mean)}
names(Av.lambda)<-c("Av.best","Av.tol.1SE","Av.tol","Av.tol.15%")

#-- Calculate MSE and calibration slope (beta)-- #

Model.best = MSE.alpha.beta(coef=Av.coef.best,data.imputed=data.imputed,
Num.imputed=Num.imputed)
Model.tol.1SE= MSE.alpha.beta(coef=Av.coef.tol.1SE,data.imputed=data.imputed,
Num.imputed=Num.imputed)
Model.tol = MSE.alpha.beta(coef=Av.coef.tol,data.imputed=data.imputed,
Num.imputed=Num.imputed)
```

```
Model.tol.15= MSE.alpha.beta(coef=Av.coef.tol.15,data.imputed=data.imputed,
Num.imputed=Num.imputed)

Av.MSE.apparent.best          <- mean(Model.best[[1]])
Av.MSE.apparent.tol.1SE       <- mean(Model.tol.1SE[[1]])
Av.MSE.apparent.tol           <- mean(Model.tol[[1]])
Av.MSE.apparent.tol.15        <- mean(Model.tol.15[[1]])
Av.Alpha.apparent.best        <- mean(Model.best[[2]])
Av.Alpha.apparent.tol.1SE     <- mean(Model.tol.1SE[[2]])
Av.Alpha.apparent.tol         <- mean(Model.tol[[2]])
Av.Alpha.apparent.tol.15      <- mean(Model.tol.15[[2]])
Av.Beta.apparent.best         <- mean(Model.best[[3]])
Av.Beta.apparent.tol.1SE            <- mean(Model.tol.1SE[[3]])
Av.Beta.apparent.tol          <- mean(Model.tol[[3]])
Av.Beta.apparent.tol.15       <- mean(Model.tol.15[[3]])

Av.coef  <-cbind(Av.coef.best=Av.coef.best,Av.coef.tol.1SE=Av.coef.tol.1SE,
Av.coef.tol=Av.coef.tol,Av.coef.tol.15=Av.coef.tol.15)
row.names(Av.coef) <-attributes(coef(Fit.Caret$finalModel))$Dimnames[[1]]

Averages<-matrix(c(Av.MSE.apparent.best,Av.MSE.apparent.tol.1SE,
Av.MSE.apparent.tol,Av.MSE.apparent.tol.15,
Av.Beta.apparent.best,Av.Beta.apparent.tol.1SE,
Av.Beta.apparent.tol,Av.Beta.apparent.tol.15,
Av.Alpha.apparent.best,Av.Alpha.apparent.tol.1SE,
Av.Alpha.apparent.tol,Av.Alpha.apparent.tol.15),3,4,
byrow=TRUE)
row.names(Averages)<-c("Av.Apparent.MSE","Av.Apparent.Beta","Av.Apparent.Alpha")
colnames(Averages)<-c("Av.best","Av.tol.1SE","Av.tol","Av.tol.15%")

out.temp <- list(Averages=Averages,Av.lambda=Av.lambda,Av.coef=Av.coef,
missforest=Imput[[1]])
out.temp
}
```

## B.3 Harrell bootstrap validation for MissForest-Lasso

```
#FUNCTIONS NEEDED:_____________________________________________________

### 1st FUNCTION:_____________________________________________________
### A function to create the bootstrap data sets ###################################
Create.bootstrap.data.MissForest <- function (k,Num.imputed, Data.NA,
cores_2_use,parallelize,seeds){
#- Sampling the incomplete data
data.imputed.Boot <- vector("list",Num.imputed)
set.seed(seeds[k])
Boot.row <- sample(nrow(Data.NA), replace=TRUE)

#- Peformimg  missForest
Imput<- vector("list",Num.imputed)
h=1
while (h <= Num.imputed){
cl <- makeCluster(cores_2_use)
clusterExport(cl, c("Data.NA","parallelize","cores_2_use","seeds","k",
"Imput"),envir=environment())
clusterSetRNGStream(cl, seeds[k])
clusterEvalQ(cl, library(missForest))
Imput[[h]] <-
parLapply(cl = cl, X = 1:cores_2_use, fun = function(no){
missForest(Data.NA, maxiter = 10, ntree = 100, variablewise = FALSE,
```

```
verbose = TRUE,
replace = TRUE,
parallelize = parallelize)
})
stopCluster(cl)
data.imputed.Boot[[h]]=Imput[[h]][[1]]$ximp
h<-h+1
}
out<-data.imputed.Boot
out
}

### 2nd FUNCTION:_____________________________________
### A function to compute optimism ################################
optimism.alpha.beta.boot <- function (coef,data.imputed.Boot,data.imputed,
Num.imputed) {
Alpha.boot <- rep(NA,Num.imputed); Beta.boot <- rep(NA,Num.imputed)
Alpha.data <- rep(NA,Num.imputed); Beta.data <- rep(NA,Num.imputed)
MSE.data <- rep(NA,Num.imputed); MSE.boot <- rep(NA,Num.imputed)
Optimism <- rep(NA,Num.imputed);
Alpha.optimism <- rep(NA,Num.imputed); Beta.optimism<-rep(NA,Num.imputed)

i=1
while(i <= Num.imputed){
Outcome.boot <- data.imputed.Boot[[i]][,dim(data.imputed.Boot[[i]])[2]]
Outcome.data <- data.imputed[[i]][,dim(data.imputed[[i]])[2]]
modelMatrix.boot <- model.Matrix(as.formula(paste(colnames(data.imputed.Boot[[i
]])[dim(data.imputed.Boot[[i]])[2]],"~.%in%",
colnames(data.imputed.Boot[[i]])[1],"+.",sep="")),
data.imputed.Boot[[i]])
X.boot<- modelMatrix.boot[,-1]
modelMatrix.data <- model.Matrix(as.formula(paste(colnames(data.imputed[[i
]])[dim(data.imputed[[i]])[2]],"~.%in%",
colnames(data.imputed[[i]])[1],"+.",sep="")),
data.imputed[[i]])
X.data<-modelMatrix.data[,-1]
Predicted.data <- as.matrix(coef[1] + X.data%*%coef[-1])
Predicted.boot <- as.matrix(coef[1] + X.boot%*%coef[-1])
MSE.boot[i] <- mean((Outcome.boot-Predicted.boot)^2)
MSE.data[i] <- mean((Outcome.data-Predicted.data)^2)
Optimism[i] <- MSE.boot[i]-MSE.data[i]
lm.coef.data <- lm(Outcome.data ~ Predicted.data)$coef
lm.coef.boot<- lm(Outcome.boot ~ Predicted.boot)$coef
Alpha.data[i] <- lm.coef.data[1]
Beta.data[i] <- lm.coef.data[2]
Alpha.boot[i] <- lm.coef.boot[1]
Beta.boot[i] <- lm.coef.boot[2]
Alpha.optimism[i] <- Alpha.boot[i]-Alpha.data[i]
Beta.optimism[i] <- Beta.boot[i]-Beta.data[i]

i<-i+1
}
out.opti.best.tol <- list(MSE.boot,MSE.data,Optimism,Alpha.optimism,
Beta.optimism)
out.opti.best.tol
}


### 3rd FUNCTION_____________________________________________
# HARREL BOOTSTRAP VALIDATION ###################################################
```

```
MissForest.Boot<- function (Num.imputed,Num.boot=NULL,method,Num.cv=NULL,
repeats=NULL,Data.NA,Grid,percent.tol,cores_2_use,parallelize,
missingData,seed){

###- Peformimg missForest
set.seed(seed)
seeds<-sample(1:10000000,100)
data.imputed <- vector("list",Num.imputed)

Imput<- vector("list",Num.imputed)
j=1
while (j <= Num.imputed){
cl <- makeCluster(cores_2_use)
clusterExport(cl, c("Data.NA","parallelize","cores_2_use","seed","Imput"),
envir=environment())
clusterSetRNGStream(cl, seed)
clusterEvalQ(cl, library(missForest))
Imput[[j]] <-
parLapply(cl = cl, X = 1:cores_2_use, fun = function(no){
missForest(Data.NA, maxiter = 10, ntree = 100, variablewise = FALSE,
verbose = TRUE,
replace = TRUE,
parallelize = parallelize)
})
stopCluster(cl)
data.imputed[[j]]=Imput[[j]][[1]]$ximp

j<-j+1
}

###- Fit glmnet (over the bootstrap data sets) using caret.
All.MSE.boot.best      <- All.MSE.data.best      <- rep(NA,Num.boot)
All.MSE.boot.tol1SE    <- All.MSE.data.tol1SE    <- rep(NA,Num.boot)
All.MSE.boot.tolerance<- All.MSE.data.tolerance<- rep(NA,Num.boot)
All.MSE.boot.tol15     <- All.MSE.data.tol15     <- rep(NA,Num.boot)
All.Optimism.best      <- All.Optimism.tol1SE    <- All.Optimism.tolerance<-
All.Optimism.tol15 <- rep(NA,Num.boot)
All.Beta.best.opt      <- All.Beta.tol1SE.opt    <- All.Alpha.best.opt     <-
All.Alpha.tol1SE.opt<- rep(NA,Num.boot)
All.Beta.tolerance.opt<- All.Alpha.tolerance.opt<- All.Beta.tol15.opt     <-
All.Alpha.tol15.opt <- rep(NA,Num.boot)

modelMatrixData<-model.Matrix(as.formula(paste(colnames(Data.NA)[dim(Data.NA)[2
]],"~.%in%",colnames(Data.NA)[1],"+.",sep="")),model.frame(
Data.NA))

Fit.caret.best <- matrix(NA,Num.imputed,dim(modelMatrixData)[2])
Fit.caret.tol1SE<- matrix(NA,Num.imputed,dim(modelMatrixData)[2])
Fit.caret.tolerance <- matrix(NA,Num.imputed,dim(modelMatrixData)[2])
Fit.caret.tol15 <- matrix(NA,Num.imputed,dim(modelMatrixData)[2])

k=1
while (k <= Num.boot){
out <-  Create.bootstrap.data.Mice.Forest.Mus(k=k,Data.NA=Data.NA,
Num.imputed=Num.imputed,cores_2_use = cores_2_use,parallelize =
parallelize,seeds=seeds)
data.imputed.Boot<- out

i=1
while (i <= Num.imputed){
modelMatrix<-model.Matrix(as.formula(paste(colnames(data.imputed.Boot[[i]])[
```

```
dim(data.imputed.Boot[[i]])[2]],"~.%in%",
colnames(data.imputed.Boot[[i]])[1],"+.",sep=""))),
data.imputed.Boot[[i]])
Outcome.boot<- data.imputed.Boot[[i]][,dim(data.imputed.Boot[[i]])[2]]
#### modified

options(warn=-1)
cl=makeCluster(cores_2_use);registerDoParallel(cl)
if (method=="repeatedcv")
{set.seed(seeds[k])
Fit.Caret <- train(modelMatrix[,-1] ,Outcome.boot, method="glmnet",
tuneGrid=Grid, family="gaussian",
trControl=trainControl(number=Num.cv, method=method,
repeats=repeats, selectionFunction="best"))
} else if(method=="boot")
{set.seed(seeds[k])
Fit.Caret <- train(modelMatrix[,-1] ,Outcome.boot, method="glmnet",
tuneGrid=Grid, family="gaussian",
trControl=trainControl(number=Num.boot,
method=method, selectionFunction="best"))
}
stopCluster(cl)
options(warn=0)

#- Model coefficients -#

Fit.caret.best[i,] <- as.matrix(coef(Fit.Caret$finalModel,s=
Fit.Caret$bestTune$lambda))
Fit.caret.tol1SE[i,] <- as.matrix(coef(Fit.Caret$finalModel,s=
max(Fit.Caret$results$lambda[Fit.Caret$results$
RMSE<=Fit.Caret$results[row.names(Fit.Caret$bestTune
),]$RMSE +(Fit.Caret$results[row.names(Fit.Caret$
bestTune),]$RMSESD)/sqrt(Num.boot)])))

Fit.caret.tolerance[i,] <- as.matrix(coef(Fit.Caret$finalModel,s=
max(Fit.Caret$results$lambda[(Fit.Caret$results$
RMSE-min(Fit.Caret$results$RMSE))*100/
min(Fit.Caret$results$RMSE)<=percent.tol])))
Fit.caret.tol15[i,]<- as.matrix(coef(Fit.Caret$finalModel,s=max(Fit.Caret$
results$lambda[(Fit.Caret$results$RMSE-min(Fit.Caret$
results$RMSE))*100/min(Fit.Caret$results$RMSE)<=15])))
i<-i+1
}

coef.best <- apply(Fit.caret.best,2,mean)
coef.tol1SE <- apply(Fit.caret.tol1SE,2,mean)
coef.tolerance <- apply(Fit.caret.tolerance,2,mean)
coef.tol15 <- apply(Fit.caret.tol15,2,mean)

#-- Calculate optimism and calibration slope (beta)-- #

Model.best = optimism.alpha.beta.deb.mus(coef=coef.best,data.imputed.Boot=
data.imputed.Boot,data.imputed=data.imputed,Num.imputed=
Num.imputed)
Model.tol.1SE = optimism.alpha.beta.deb.mus(coef=coef.tol1SE,
data.imputed.Boot=data.imputed.Boot,data.imputed=data.imputed,
Num.imputed=Num.imputed)
Model.tol = optimism.alpha.beta.deb.mus(coef=coef.tolerance,data.imputed.Boot=
data.imputed.Boot,data.imputed=data.imputed,Num.imputed=Num.imputed)
Model.tol.15 = optimism.alpha.beta.deb.mus(coef=coef.tol15,data.imputed.Boot=
data.imputed.Boot,data.imputed=data.imputed,Num.imputed=Num.imputed)
```

```
All.MSE.boot.best[k]      <- mean(Model.best[[1]])
All.MSE.data.best[k]      <- mean(Model.best[[2]])
All.MSE.boot.tol1SE[k]    <- mean(Model.tol.1SE[[1]]) #training
All.MSE.data.tol1SE[k]    <- mean(Model.tol.1SE[[2]]) #test
All.MSE.boot.tolerance[k]<- mean(Model.tol[[1]]) #training
All.MSE.data.tolerance[k]<- mean(Model.tol[[2]]) #test
All.MSE.boot.tol15[k]     <- mean(Model.tol.15[[1]]) #training
All.MSE.data.tol15[k]     <- mean(Model.tol.15[[2]]) #test

All.Optimism.best[k]     <- mean(Model.best[[3]])
All.Alpha.best.opt[k]    <- mean(Model.best[[4]])
All.Beta.best.opt[k]     <- mean(Model.best[[5]])
All.Optimism.tol1SE[k]   <- mean(Model.tol.1SE[[3]])
All.Alpha.tol1SE.opt[k] <- mean(Model.tol.1SE[[4]])
All.Beta.tol1SE.opt[k]  <- mean(Model.tol.1SE[[5]])
All.Optimism.tolerance[k]<- mean(Model.tol[[3]])
All.Alpha.tolerance.opt[k]<- mean(Model.tol[[4]])
All.Beta.tolerance.opt[k] <- mean(Model.tol[[5]])
All.Optimism.tol15[k]    <- mean(Model.tol.15[[3]])
All.Alpha.tol15.opt[k]   <- mean(Model.tol.15[[4]])
All.Beta.tol15.opt[k]    <- mean(Model.tol.15[[5]])

k<-k+1
}
## Calcualte MSE, optimism and model averaged regression coefficents
Av.MSE.boot.best       <- mean(All.MSE.boot.best) #training
Av.MSE.data.best       <- mean(All.MSE.data.best) #test
Av.MSE.boot.tol1SE     <- mean(All.MSE.boot.tol1SE) #training
Av.MSE.data.tol1SE     <- mean(All.MSE.data.tol1SE) #test
Av.MSE.boot.tolerance <- mean(All.MSE.boot.tolerance) #training
Av.MSE.data.tolerance <- mean(All.MSE.data.tolerance) #test
Av.MSE.boot.tol15      <- mean(All.MSE.boot.tol15) #training
Av.MSE.data.tol15      <- mean(All.MSE.data.tol15) #test
Av.Optimism.best       <- mean(All.Optimism.best)
Av.Optimism.tol1SE     <- mean(All.Optimism.tol1SE)
Av.Optimism.tolerance <- mean(All.Optimism.tolerance)
Av.Optimism.tol15      <- mean(All.Optimism.tol15)
Av.Beta.best.opt       <- mean(All.Beta.best.opt)
Av.Alpha.best.opt      <- mean(All.Alpha.best.opt)
Av.Beta.tol1SE.opt     <- mean(All.Beta.tol1SE.opt)
Av.Alpha.tol1SE.opt    <- mean(All.Alpha.tol1SE.opt)
Av.Beta.tolerance.opt <- mean(All.Beta.tolerance.opt)
Av.Alpha.tolerance.opt<- mean(All.Alpha.tolerance.opt)
Av.Beta.tol15.opt      <- mean(All.Beta.tol15.opt)
Av.Alpha.tol15.opt     <- mean(All.Alpha.tol15.opt)

Averages<-matrix(c(Av.MSE.boot.best,Av.MSE.boot.tol1SE,Av.MSE.boot.tolerance,
Av.MSE.boot.tol15,
Av.MSE.data.best,
Av.MSE.data.tol1SE,Av.MSE.data.tolerance,Av.MSE.data.tol15,
Av.Optimism.best,Av.Optimism.tol1SE,Av.Optimism.tolerance,
Av.Optimism.tol15,
Av.Beta.best.opt,Av.Beta.tol1SE.opt,Av.Beta.tolerance.opt,
Av.Beta.tol15.opt,
Av.Alpha.best.opt, Av.Alpha.tol1SE.opt,
Av.Alpha.tolerance.opt,Av.Alpha.tol15.opt),5,4,byrow=TRUE)
colnames(Averages)<-c("Av.best","Av.tol.1SE","Av.tol","Av.tol.15%")
# }
row.names(Averages)<-c("Av.Training.MSE","Av.Test.MSE","Av.Optimism",
```

```
"Av.Beta.optimism","Av.Alpha.optimism")

out.temp <- list(Averages=Averages)
out.temp
}
```

# Appendix C

# Database of cognitive training and remediation studies

## C.1 Study information variables

**Database of Cognitive Training and Remediation Studies - Study Information**

| | | |
|---|---|---|
| field 1 = | Study ID number (assigned by NIMH)<br>4155 = WYKES1 (n=35)<br>5693 = BELL (n=77)<br>6632 = KEEFE (n=53)<br>7926 = KESHAVAN (n=58)<br>8134 = WYKES3 (n=40)<br>9212 = WYKES2 (n=85)<br>9479 = SILVERSTEIN (n=83) | (Study_ID) |
| field 2 = | PI last name | (PI_Last_Name) |

**SECTION 1: General study information**

Note: The following 4 fields refer to the primary paper in which the results of the study are reported; If the results are reported in multiple papers, generally information about the initial report of results is included here.

| | | |
|---|---|---|
| field 3 = | First author of paper (Last name) | (Paper_1stAuthor) |
| field 4 = | Title of paper | (Paper_Title) |
| field 5 = | Journal name | (Paper_Journal) |
| field 6 = | Publication date (year) | (Paper_Pubyear) |

**SECTION 2: Summary of subject and study characteristics**

Note: Data for the variables in this section are computed from the subject-level data for each study.

| | | |
|---|---|---|
| field 7 = | Number of subjects in Cognitive Remediation condition | (N_CR_Group) |
| field 8 = | Number of subjects in Comparison condition | (N_Comp_Group) |
| field 9 = | Percentage of participants who are male | (Percent_Male) |
| field 10 = | Mean age of participants (years) | (Mean_Age) |
| field 11 = | Percentage of participants who are not White | (Percent_Not_White) |

Note: Values for the following variables are obtained via the Study Information form completed by the PI.

| | | |
|---|---|---|
| field 12 = | Data collection start date (year) | (Study_Start_Date) |
| field 13 = | Data collection end date (year) | (Study_End_Date) |
| field 14 = | Location of data collection; Country 1 | (Study_Country1) |
| field 15 = | Location of data collection; Country 2 | (Study_Country2) |
| field 16 = | Location of data collection; Country 3 | (Study_Country3) |

10 = Canada
20 = France
30 = Germany
40 = Italy
50 = Mexico
60 = Netherlands
70 = Spain
80 = United States
90 = United Kingdom

## Database of Cognitive Training and Remediation Studies - Study Information

**SECTION 3: Intervention and comparison condition characteristics**

Note: Both experimental conditions may include CR but "CR condition" is defined (per the study information form) as being the most elaborated intervention

Codes for following variables:
1=Yes
0=No

| | | |
|---|---|---|
| field 17 = | Attention shaping intervention used in CR condition? | (Int_CR_Attn) |
| field 18 = | Attention shaping intervention used in comparison condition? | (Int_Con_Attn) |
| field 19 = | Cognitive Enhancement Tx used in CR condition? | (Int_CR_CET) |
| field 20 = | Cognitive Enhancement Tx used in comparison condition? | (Int_Con_CET) |
| field 21 = | Cognitive Remediation Therapy used in CR condition? | (Int_CR_CRT) |
| field 22 = | Cognitive Remediation Therapy used in comparison condition? | (Int_Con_CRT) |
| field 23 = | Integrated Psychological Therapy used in CR condition? | (Int_CR_IPT) |
| field 24 = | Integrated Psychological Therapy used in comparison condition? | (Int_Con_IPT) |
| field 25 = | Neuropsychological Educational Approach to Rehabilitation used in CR condition? | (Int_CR_NEAR) |
| field 26 = | Neuropsychological Educational Approach to Rehabilitation used in comparison condition? | (Int_Con_NEAR) |
| field 27 = | Neurocognitive Enhancement Therapy used in CR condition? | (Int_CR_NET) |
| field 28 = | Neurocognitive Enhancement Therapy used in comparison condition? | (Int_Con_NET) |

Text field:

| | | |
|---|---|---|
| field 29 = | Name of other CR approach used | (OtherCRname) |

Codes for following variables:
1=Yes
0=No

| | | |
|---|---|---|
| field 30 = | Other CR approach used in CR condition? | (Int_CR_Other) |
| field 31 = | Other CR approach used in comparison condition? | (Int_Con_Other) |
| field 32 = | Non-remediation computer tasks/games used in CR condition? | (Int_CR_Games) |
| field 33 = | Non-remediation computer tasks/games used in comparison condition? | (Int_Con_Games) |
| field 34 = | Cognitive-Behavioral therapy used in CR condition? | (Int_CR_CBT) |
| field 35 = | Cognitive-Behavioral therapy used in comparison condition? | (Int_Con_CBT) |

Text field:

| | | |
|---|---|---|
| field 36 = | Name of cognition-enhancing medication | (Cog_Med_Name) |

**Database of Cognitive Training and Remediation Studies - Study Information**

Codes for following variables:
1=Yes
0=No

field 37 = Putative cognition-enhancing medication used in CR condition? (Int_CR_Med)
field 38 = Putative cognition-enhancing medication used in comparison condition? (Int_Con_Med)
field 39 = Social skills training used in CR condition? (Int_CR_SST)
field 40 = Social skills training used in comparison condition? (Int_Con_SST)
field 41 = Supported employment used in CR condition? (Int_CR_SuppEmp)
field 42 = Supported employment used in comparison condition? (Int_Con_SuppEmp)
field 43 = Other voc rehabilitation used in CR condition? (Int_CR_Voc)
field 44 = Other voc rehabilitation used in comparison condition? (Int_Con_Voc)
field 45 = General psychosocial rehabilitation used in CR condition? (Int_CR_Rehab)
field 46 = General psychosocial rehabilitation used in comparison condition? (Int_Con_Rehab)

Text field:

field 47 = Name of first other intervention used in CR or Comparison condition (Other_intervention1)

Codes for following variables:
1=Yes
0=No

field 48 = Other intervention used in CR condition? (Int_CR_Other1)
field 49 = Other intervention used in comparison condition? (Int_Con_Other1)

Text field:

field 50 = Name of second other intervention used in CR or Comparison condition (Other_intervention2)

Codes for following variables:
1=Yes
0=No

field 51 = Second other intervention used in CR condition? (Int_CR_Other2)
field 52 = Second other intervention used in comparison condition? (Int_Con_Other2)
field 53 = Patients in CR condition get psych meds/med management? (Int_CR_MedMgt)
field 54 = Patients in Comparison condition get psych meds/med management? (Int_Con_MedMgt)

Text field:

field 55 = Description of additional experimental condition, if more than 2 study groups (Other_condition)

## Database of Cognitive Training and Remediation Studies - Study Information

**SECTION 4: Techniques used in cognitive remediation intervention**

Codes for following variables:
1=Most central to the intervention
2=Second most central
3=Third most central
4=Fourth most central

| | | |
|---|---|---|
| field 56 = | Rank order of Drill and Practice technique | (Techn_DandP) |
| field 57 = | Rank order of Strategy training technique | (Techn_Strategy) |
| field 58 = | Rank order of metacognitive training technique | (Techn_MetaCog) |
| field 59 = | Rank order of errorless learning technique | (Techn_Errorless) |
| field 60 = | Rank order of social praise technique | (Techn_Social_Praise) |
| field 61 = | Rank order of tangible rewards technique | (Techn_Rewards) |
| field 62 = | Rank order of rehearsal technique | (Techn_Rehearsal) |
| field 63 = | Rank order of habit training technique | (Techn_Habit) |
| field 64 = | Rank order of compensatory techniques | (Techn_Compensatory) |
| field 65 = | Rank order of group processing technique | (Techn_Group) |
| field 66 = | Rank order of in vivo practice techniques | (Techn_InVivo) |
| field 67 = | Rank order of other technique | (Techn_Other) |

Text field:

| | | |
|---|---|---|
| field 68 = | Description of other CR technique | (Describe_OtherTechn) |

**SECTION 5: Treatment targets**

Codes for following variables:
1=Primary target of CR intervention
2=Secondary target
3=Tertiary target
4=Fourth priority target

| | | |
|---|---|---|
| field 69 = | Rank order of general cognition among targets of CR intervention | (Target_Cognition) |
| field 70 = | Rank order of skill acquisition among targets of CR intervention | (Target_SkillAcq) |
| field 71 = | Rank order of work functioning among targets of CR intervention | (Target_Work) |
| field 72 = | Rank order of attention among targets of CR intervention | (Target_Attention) |
| field 73 = | Rank order of verbal memory among targets of CR intervention | (Target_VerbalMem) |
| field 74 = | Rank order of social cognition among targets of CR intervention | (Target_SocialCog) |
| field 75 = | Rank order of social skills among targets of CR intervention | (Target_SocialSkills) |
| field 76 = | Rank order of other target among targets of CR intervention | (Target_Other) |

Text field:

| | | |
|---|---|---|
| field 77 = | Description of other target of CR intervention | (Describe_Other_Target) |

Codes for following variable:

1=Primary target of CR intervention
2=Secondary target
3=Tertiary target
4=Fourth priority target

field 78 = Rank order of additional other target among targets of CR intervention (Target_Other2)

Text field:

field 79 = Description of additional other target of CR intervention (Describe_Other2_Target)
field 80 = Comments about study conditions and CR intervention (Comments_Conditions_CR)

**SECTION 6: Cognitive remediation delivery methods**

Codes for following variables:
1=Yes
0=No

field 81 = CR computerized? (Delivery_Computerized)
field 82 = CR delivered via paper and pencil? (Delivery_PaperPencil)
field 83 = CR delivered via mix of computerized and paper/pencil? (Delivery_Mixed)
field 84 = Other method of delivering CR used? (Delivery_Other)

Text field:

field 85 = Description of other method of delivering CR (Describe_Delivery_Other)

**SECTION 7: Cognitive remediation delivery format**

Codes for following variables:
1=Yes
0=No

field 86 = CR sessions delivered one-on-one? (Format_One)
field 87 = CR sessions delivered in group format? (Format_Group)
field 88 = CR sessions delivered in a mix of individual and group sessions? (Format_Mix)
field 89 = CR sessions delivered in another format? (Format_Other)

Text field:

field 90 = Description of other format in which CR was delivered (Describe_Format_Other)

## Database of Cognitive Training and Remediation Studies - Study Information

**SECTION 8: Cognitive remediation delivery setting**

Codes for following variables:
1=Yes
0=No

| | | |
|---|---|---|
| field 91 = | CR delivered in an outpatient psychiatry/ mental health clinic? | (Setting_Outpt) |
| field 92 = | CR delivered in a day treatment center? | (Setting_DayTx) |
| field 93 = | CR delivered on an inpatient unit? | (Setting_Inpt) |
| field 94 = | CR delivered in home setting? | (Setting_Home) |
| field 95 = | CR delivered in other setting? | (Setting_Other) |

Text field:

| | | |
|---|---|---|
| field 96 = | Description of other setting where CR delivered | (Describe_Setting_Other) |

Codes for following variables:
1=Yes
0=No

| | | |
|---|---|---|
| field 97 = | Did CR sessions take place in > 1 setting per subject? | (Multi_Setting) |
| field 98 = | Did CR sessions take place in different settings for different subjects? | (Multi_Site) |

**SECTION 9: Other treatment features**

Entered responses for the following variables:

| | | |
|---|---|---|
| field 99 = | Typical duration of CR session | (Session_Duration) |
| field 100 = | Target # of CR sessions per week | (Session_Freq) |
| field 101 = | Target duration of CR intervention (in weeks) | (CR_Duration) |

Codes for following variable:
1=Yes
0=No

| | | |
|---|---|---|
| field 102 = | Duration and frequency of control condition similar to CR? | (Con_Freq_Dur) |

Text field:

| | | |
|---|---|---|
| field 103 = | If frequency and duration of control sessions different than CR, describe | (Con_FreqDur_Desc) |

**Database of Cognitive Training and Remediation Studies - Study Information**

Codes for following variables:
1=Yes
0=No

field 104 = Did doctoral-level clinicians administer CR? (Clinicians_Doctoral)
field 105 = Did masters-level clinicians administer CR? (Clinicians_Masters)
field 106 = Did trainers w/out graduate training administer CR? (Clinicians_NoGrad)
field 107 = Were CR trainers research staff? (Research_Staff)
field 108 = Were CR trainers clinical staff? (Clinical_Staff)
field 109 = Subject paid for CR and assessment sessions? (Pay_All)

Entered response for the following variable:

field 110 = Amount paid for each CR/Con session (dollars) (Pay_Session)

Codes for following variables:
1=Yes
0=No

field 111 = Subjects paid for assessments only? (Pay_AssessOnly)
field 112 = Subjects compensated non-monetarily for CR/Control sessions? (Pay_Nonmon)
field 113 = Subjects in CR compensated for sessions but control subjects not compensated? (Pay_CR_only)

Text field:

field 114 = Comments about delivery of CR/Con intervention (Comments_tx_delivery)

**SECTION 10: Defining "completion"**

Text field:

field 115 = Criteria for identifying participants as "completers" (Completers_Criteria)

Codes for following variables:
1=Yes
0=No

field 116 = Are data from all participants who were randomized into the study included in database? (All_participants)
field 117 = Are data only from "completers" included in database? (Completers_Only)
field 118 = Other data are included in database (Completers_Other)

Text field:

field 119 = Comment about which subjects included in database (Completers_Other_Comment)

**SECTION 11: Randomization, blinding and adherence**

Codes for following variables:
1=Yes
0=No

field 120 = Randomization (or minimization allocation) procedure used? (Randomization_Used)
field 121 = Was randomization carried out independently from the study team? (Randomization_Indep)
field 122 = Study staff who were not CR therapists did assessments? (Assessors_StaffNonTx)
field 123 = Study staff who were CR therapists did assessments? (Assessors_StaffTx)
field 124 = Non-study staff did assessments? (Assessors_Nonstaff)
field 125 = Were assessors blind/masked to participant's treatment group assignment? (Assessors_Blind)
field 126 = Was assessors' blinding verified? (Assessors_Blindchecked)
field 127 = Was a treatment manual or protocol used? (Tx_Manualized)
field 128 = Was adherence to treatment protocol and quality of treatment delivery assessed on an ongoing basis? (Adherence_Checked)

Text field:

field 129 = Comments about randomization, blinding, assessors and adherence/quality assessment (Randomization_Comments)

**SECTION 12: Eligibility criteria**

Note: These data are obtained from the list of inclusion/exclusion criteria provided by PIs; there is no standard reporting form and not all studies will include all of these criteria

Entered responses for the following variables:

field 130 = Minimum age for eligibility (in years) (Elig_Min_age)
field 131 = Maximum age for eligibility (in years) (Elig_Max_age)

Codes for following variables:
1=Yes
0=No

field 132 = Patients with diagnosis of schizophrenia eligible? (Elig_Dx_Schiz)
field 133 = Patients with diagnosis of schizoaffective disorder eligible? (Elig_Dx_Schizoaffective)

Codes for following variable:
1 = Schizophreniform
2 = Bipolar disorder

field 134 = If patients with diagnoses other than schizophrenia or schizoaffective were eligible for study, which other diagnoses were included? (Elig_DX_Other)

Text fields:

| | | |
|---|---|---|
| field 135 = | If med stability is an inclusion criterion, what is the criterion? | (Elig_MedStability) |
| field 136 = | If medication type is an exclusionary criterion, what types of medications were prohibited? | (Elig_MedType) |
| field 137 = | If symptom severity an eligibility criterion, what is the criteria? | (Elig_Sx) |
| field 138 = | If English proficiency is an inclusion criterion, how was proficiency defined? | (Elig_English) |
| field 139 = | If IQ is an inclusion criterion, what was cutoff? | (Elig_IQ) |
| field 140 = | If cognitive impairment is an inclusion criterion, what is the criterion? | (Elig_Cog_Impair) |
| field 141 = | If social functioning impairment was an inclusion criterion, what was the criterion? | (Elig_SocialFx) |
| field 142 = | If mental retardation, pervasive developmental disorder, and/or neurological disorders were exclusionary, what was criterion? | (Elig_Neuro) |
| field 143 = | If substance use disorder is an exclusionary criterion, what was the criterion? | (Elig_SUD) |

Codes for following variables:
1=Yes
0=No

| | | |
|---|---|---|
| field 144 = | Prior cognitive remediation an exclusionary criterion? | (Elig_PriorCR) |

Text fields:

| | | |
|---|---|---|
| field 145 = | Other inclusion/exclusion criterion 1 | (Elig_Other1) |
| field 146 = | Other inclusion/exclusion criterion 2 | (Elig_Other2) |
| field 147 = | Other inclusion/exclusion criterion 3 | (Elig_Other3) |
| field 148 = | Other inclusion/exclusion criterion 4 | (Elig_Other4) |

**SECTION 13: Summary of study assessment schedule**

Note: PIs were asked to complete these items on the Study Information form only if they did not provide a detailed schedule of assessments.

Entered responses for the following variables:

| | | |
|---|---|---|
| field 149 = | Interval between baseline and 1st midpoint assessment (in weeks) | (Mid_Point1) |
| field 150 = | Interval between 1st mid-point and 2nd midpoint assessment (in weeks) | (Mid_Point2) |
| field 151 = | Interval between baseline and end-of-treatment assessment (in weeks) | (Post_Treatment) |
| field 152 = | Interval between post-treatment and 1st follow-up assessment (in weeks) | (Followup_1) |
| field 153 = | Interval between 1st follow-up assessment and 2nd follow-up assessment (in weeks) | (Followup_2) |
| field 154 = | Interval between 2nd follow-up and 3rd follow-up assessment (in weeks) | (Followup_3) |

**Database of Cognitive Training and Remediation Studies - Study Information**

Codes for following variables:
1=Yes
0=No

| | | |
|---|---|---|
| field 155 = | Cognitive measures done at baseline assessment? | (Cog_Baseline) |
| field 156 = | Cognitive measures done at 1st midpoint assessment? | (Cog_Midpoint1) |
| field 157 = | Cognitive measures done at 2nd midpoint assessment? | (Cog_Midpoint2) |
| field 158 = | Cognitive measures done at post-treatment assessment? | (Cog_PostTx) |
| field 159 = | Cognitive measures done at 1st follow-up assessment? | (Cog_Followup1) |
| field 160 = | Cognitive measures done at 2nd follow-up assessment? | (Cog_Followup2) |
| field 161 = | Cognitive measures done at 3rd follow-up assessment? | (Cog_Followup3) |
| field 162 = | Symptom measures done at baseline assessment? | (Sx_Baseline) |
| field 163 = | Symptom measures done at 1st midpoint assessment? | (Sx_Midpoint1) |
| field 164 = | Symptom measures done at 2nd midpoint assessment? | (Sx_Midpoint2) |
| field 165 = | Symptom measures done at post- treatment assessment? | (Sx_PostTx) |
| field 166 = | Symptom measures done at 1st follow-up assessment? | (Sx_Followup1) |
| field 167 = | Symptom measures done at 2nd follow-up assessment? | (Sx_Followup2) |
| field 168 = | Symptom measures done at 3rd follow-up assessment? | (Sx_Followup3) |
| field 169 = | Functioning measures done at baseline assessment? | (Fx_Baseline) |
| field 170 = | Functioning measures done at 1st midpoint assessment? | (Fx_Midpoint1) |
| field 171 = | Functioning measures done at 2nd midpoint assessment? | (Fx_Midpoint2) |
| field 172 = | Functioning measures done at post- treatment assessment? | (Fx_PostTx) |
| field 173 = | Functioning measures done at 1st follow-up assessment? | (Fx_Followup1) |
| field 174 = | Functioning measures done at 2nd follow-up assessment? | (Fx_Followup2) |
| field 175 = | Functioning measures done at 3rd follow-up assessment? | (Fx_Followup3) |

Text field:

| | | |
|---|---|---|
| field 176 = | Description of other measure | (Other_Measure) |

Codes for following variables:
1=Yes
0=No

| | | |
|---|---|---|
| field 177 = | Other measures done at baseline assessment? | (Other_Baseline) |
| field 178 = | Other measures done at 1st midpoint assessment? | (Other_Midpoint1) |
| field 179 = | Other measures done at 2nd midpoint assessment? | (Other_Midpoint2) |
| field 180 = | Other measures done at post- treatment assessment? | (Other_PostTx) |
| field 181 = | Other measures done at 1st follow-up assessment? | (Other_Followup1) |
| field 182 = | Other measures done at 2nd follow-up assessment? | (Other_Followup2) |
| field 183 = | Other measures done at 3rd follow-up assessment? | (Other_Followup3) |

Text field:

| | | |
|---|---|---|
| field 184 = | Description of other2 measure | (Other_Measure2) |

Codes for following variables:

1=Yes
0=No

| | | |
|---|---|---|
| field 185 = | Other2 measures done at baseline assessment? | (Other2_Baseline) |
| field 186 = | Other2 measures done at 1st midpoint assessment? | (Other2_Midpoint1) |
| field 187 = | Other2 measures done at 2nd midpoint assessment? | (Other2_Midpoint2) |
| field 188 = | Other2 measures done at post- treatment assessment? | (Other2_PostTx) |
| field 189 = | Other2 measures done at 1st follow-up assessment? | (Other2_Followup1) |
| field 190 = | Other2 measures done at 2nd follow-up assessment? | (Other2_Followup2) |
| field 191 = | Other2 measures done at 3rd follow-up assessment? | (Other2_Followup3) |

Text field:

| | | |
|---|---|---|
| field 192 = | Comments re: schedule of assessments | (Comments_Schedule) |

# C.2 Cognitive variables

## Database of Cognitive Training and Remediation Studies–Cognitive Data (COG)

field 1 = Study ID number (assigned by NIMH) (Study_ID)
- 4155 = WYKES1 (n=35)
- 5693 = BELL (n=77)
- 6632 = KEEFE (n=53)
- 7926 = KESHAVAN (n=58)
- 8134 = WYKES3 (n=40)
- 9212 = WYKES2 (n=85)
- 9479 = SILVERSTEIN (n=83)

field 2 = Participant ID number (assigned by NIMH) (Participant_ID)

field 3 = Study condition to which participant was assigned (Study_Condition)
- 1 = Cognitive remediation
- 2 = Comparison condition
- 3 = Other

field 4 = Assessment time point (Time_point)
- 5 = Screening
- 10 = Baseline
- 21, 22, 23, … = Mid-point 1, 2, 3, …
- 30 = End-of-treatment
- 41, 42, 43, … = Follow-up1, 2, 3, …

field 5 = Days since baseline (Days_Baseline)

**NOTE: The Days_Baseline variable is calculated using assessment dates when they were provided and calculated from target dates according to the schedule of assessment time points for the study when visits dates were not provided.**

field 6 = Norms used for tests included in the MCCB (Norms_MCCB)
- 1 = Original norms provided by test publisher
- 2 = MCCB norms
- 3 = Other

**NOTE: Tests included in the MATRICS Consensus Cognitive Battery (MCCB) are marked with an asterisk.**

field 7 = Ammons Quick Test Full Scale IQ (AmmQT_IQ)

field 8 = Brief Assessment of Cognition in Schizophrenia: Symbol-Coding subtest* (# correct; raw score) (BACS_SC_Raw)
field 9 = Brief Assessment of Cognition in Schizophrenia: Symbol-Coding subtest* (# correct; T-score) (BACS_SC_Tscore)
field 10 = Brief Assessment of Cognition in Schizophrenia: Digit Sequencing Task (number of correct responses; 0 to 28) (BACS_DigitSeq_NumCorrect)
field 11 = Brief Assessment of Cognition in Schizophrenia: Tower of London Task (number of correct responses; 0 to 22) (BACS_Tower_NumCorrect)

field 12 = Brief Visuospatial Memory Test – Revised* (3-trial total recall; raw score) (BVMTR_Raw)
field 13 = Brief Visuospatial Memory Test – Revised* (3-trial total recall; T-score) (BVMTR_Tscore)

field 14 = Category fluency: Animal naming* (# animals named in 60 s; raw score) (CATFLU_Raw)
field 15 = Category fluency: Animal naming* (# animals named in 60 s; T-score) (CATFLU_Tscore)

field 16 = Continuous Performance Test – Identical Pairs* (Mean d' across 2-, 3-, and 4-digit conditions; raw score) (CPT_IP_Raw)
field 17 = Continuous Performance Test – Identical Pairs* (Mean d' across 2-, 3-, and 4-digit conditions T score) (CPT_IP_Tscore)

field 18 = California Verbal Learning Test: Total Recall (CVLT_Total_Recall)
field 19 = California Verbal Learning Test: Short-term Free Recall (CVLT_Short_Freerecall)
field 20 = California Verbal Learning Test: Long-term Free Recall (CVLT_Long_Freerecall)

field 21 = Verbal fluency (FAS): Total number of correct responses (FAS_N_Responses)
field 22 = Verbal fluency (FAS): Age- and education-adjusted score (FAS_Adj_Score)

field 23 = Hopkins Verbal Learning Test – Revised* (Total # words recalled over 3 trials; raw score) (HVLTR_Raw)
field 24 = Hopkins Verbal Learning Test – Revised* (Total # words recalled over 3 trials; T-score)(HVLTR_Tscore)

field 25 = Letter-Number Span* (# of correct trials; raw score) (LNS_Raw)
field 26 = Letter-Number Span* (# of correct trials; T-score) (LNS_Tscore)

field 27 = MATRICS Speed of Processing domain* (T-Score) (MCCB_Speed_Tscore)
field 28 = MATRICS Attention/Vigilance domain* (T-Score) (MCCB_AttnVig_Tscore)
field 29 = MATRICS Working Memory domain* (T-Score) (MCCB_WorkMem_Tscore)
field 30 = MATRICS Verbal Learning domain* (T-Score) (MCCB_VerbLearn_Tscore)
field 31 = MATRICS Visual Learning domain* (T-Score) (MCCB_VisLearn_Tscore)
field 32 = MATRICS Reasoning and Problem Solving domain* (T-Score) (MCCB_ReasProb_Tscore)
field 33 = MATRICS Social Cognition domain* (T-Score) (MCCB_SocCog_Tscore)
field 34 = MATRICS Overall Composite* (T-Score) (MCCB_Overall_Tscore)

field 35 = Mayer-Salovey-Caruso Emotional Intelligence Test: Managing Emotions subtest* (raw score) (MSCEIT_MEmo_Raw)
field 36 = Mayer-Salovey-Caruso Emotional Intelligence Test: Managing Emotions subtest* (T score) (MSCEIT_MEmo_Tscore)

field 37 = Modified six elements task: Number of tasks attempted (MSET_Attempted)
field 38 = Modified six elements task: Number of rules broken (MSET_Rules)
field 39 = Modified six elements task: Total score: no. of tasks attempted minus no. of rule breaks (MSET_Total)

field 40 = Neuropsychological Assessment Battery: Mazes subtest* (Total; raw score) (NAB_Mazes_Raw)
field 41 = Neuropsychological Assessment Battery: Mazes subtest* (Total; T-score) (NAB_Mazes_Tscore)

field 42 = National Adult Reading Test : Predicted full-scale IQ (NART_FSIQEST)
field 43 = National Adult Reading Test : Predicted performance IQ (NART_PIQEST)
field 44 = National Adult Reading Test : Predicted verbal IQ (NART_VIQEST)

field 45 = Tower of London - DX: Total move score (range: 0 to 189) (TOLDX_Move)
field 46 = Tower of London - DX: Ratio of initiation to Execution time (range: 0 to 1) (TOLDX_InitExec)

field 47 = Trailmaking test part A* (Paper & pencil): Time to completion (Seconds) (TMTA_Raw)
field 48 = Trailmaking test part A* (Paper & pencil): Time to completion (T-score) (TMTA_Tscore)
field 49 = Trailmaking test Part A (Paper & pencil): Number of errors (TMTA_Errors)
field 50 = Trailmaking test Part B (Paper & pencil): Time to completion (Seconds) (TMTB_Raw)
field 51 = Trailmaking test Part B (Paper & pencil): Time to completion (T-score) (TMTB_Tscore)
field 52 = Trailmaking test Part B (Paper & pencil): Number of errors (TMTB_Errors)
field 53 = Trailmaking test Condition 1/letters (Computerized): Trial 1, Time to completion (Seconds) (TMTA_Comp_Raw)
field 54 = Trailmaking test Condition 2/letters+numbers (Computerized): Trial 1, Time to completion (Seconds) (TMTB_Comp_Raw)

field 55 = Version of Wechsler Adult Intelligence Scale used (WAIS_Version)
1 = WAIS-III
2 = WAIS-IV
3 = WAIS-R
4 = Wechsler Abbreviated Scale of Intelligence

field 56 = Wechsler Adult Intelligence Scale Picture Arrangement: Raw score (WAIS_PictArr_Raw)
field 57 = Wechsler Adult Intelligence Scale Picture Arrangement: Scaled score (WAIS_PictArr_Scaled)
field 58 = Wechsler Adult Intelligence Scale Digit-Symbol Substitution: Raw score (WAIS_DigSym_Raw)
field 59 = Wechsler Adult Intelligence Scale Digit-Symbol Substitution: Scaled score (WAIS_DigSym_Scaled)
field 60 = Wechsler Adult Intelligence Scale Digit Span: Raw score (WAIS_DigSpan_Raw)
field 61 = Wechsler Adult Intelligence Scale Digit Span: Scaled score (WAIS_DigSpan_Scaled)
field 62 = Wechsler Adult Intelligence Scale Picture Completion: Raw score (WAIS_PictComp_Raw)
field 63 = Wechsler Adult Intelligence Scale Picture Completion: Scaled score (WAIS_PictComp_Scaled)
field 64 = Wechsler Adult Intelligence Scale Vocabulary: Raw score (WAIS_Vocab _Raw)
field 65 = Wechsler Adult Intelligence Scale Vocabulary: Scaled score (WAIS_Vocab _Scaled)
field 66 = Wechsler Adult Intelligence Scale: Verbal IQ (scaled) (WAIS_VerbalIQ)
field 67 = Wechsler Adult Intelligence Scale: Performance IQ (scaled) (WAIS_PerfIQ)
field 68 = Wechsler Adult Intelligence Scale: Full-scale IQ (WAIS_FSIQ)

field 69 = Wisconsin Card Scoring Test: Administration method (WCST_method)
1 = Cards
2 = Computer

field 70 = Wisconsin Card Sorting Test: Percent Conceptual Responses (0 to 100) (WCST_Percent_Conceptual)
field 71 = Wisconsin Card Sorting Test: Percent Conceptual Reponses (Scaled) (WCST_Percent_Conceptual_Scaled)
field 72 = Wisconsin Card Sorting Test: Categories Achieved (0 to 6) (WCST_Categories)
field 73 = Wisconsin Card Sorting Test: Non-Perseverative Errors (0 to 128) (WCST_NonPersev_Errors)
field 74 = Wisconsin Card Sorting Test: Perseverative Errors (0 to 128) (WCST_Persev_Errors)

field 75 = Wechsler Memory Scale –III: Spatial Span subtest* (Sum of scores on backwards and forwards conditions; raw score) (WMS_SS_Raw)
field 76 = Wechsler Memory Scale –III: Spatial Span subtest* (Sum of scores on backwards and forwards conditions T-score) (WMS_SS_Tscore)

field 77 = Wide-Range Achievement Test: Word Reading Total (WRAT_Reading_Total)

# C.3 Demographics

## Database of Cognitive Training and Remediation– Demographics (DEMO)

| | | |
|---|---|---|
| field 1 = | **Study ID** number (assigned by NIMH)<br>4155 = WYKES1 (n=35)<br>5693 = BELL (n=77)<br>6632 = KEEFE (n=53)<br>7926 = KESHAVAN (n=58)<br>8134 = WYKES3 (n=40)<br>9212 = WYKES2 (n=85)<br>9479 = SILVERSTEIN (n=83) | (Study_ID) |
| field 2 = | **Participant ID** number (assigned by NIMH) | (Participant_ID) |
| field 3 = | **Study condition** to which participant was assigned<br>1 = Cognitive Remediation condition<br>2 = Comparison condition<br>3 = Other | (Study_condition) |
| field 4 = | **Method** for calculating days since baseline<br>1 = calculated from visit dates when provided in data<br>2 = calculated from target dates when visit dates not provided<br>3 = missing visit | (Day_Cat) |
| field 5 = | **Age** at study baseline assessment | (Pt_Baseline_Age) |
| field 6 = | **Gender**<br>1 = Female<br>2 = Male<br>3 = Unspecified<br>4 = Unknown/Missing | (Pt_Gender) |
| field 7 = | **Ethnicity**: Hispanic or Latino origin<br>1 = Yes<br>2 = No<br>3 = Unknown/Missing | (Pt_Hispanic) |
| field 8 = | **Racial** category<br>1 = American Indian or Alaskan Native<br>2 = Asian<br>3 = Black or African American<br>4 = Native Hawaiian or other Pacific Islander<br>5 = White<br>6 = Multi-racial<br>7 = Not Reported or Declined to Specify<br>8 = Other | (Pt_Race) |
| field 9 = | **Original Study Text: Race** | (Pt_Race_orig) |

## Database of Cognitive Training and Remediation– Demographics (DEMO)

Codes for **Marital Status** variables:

1 = Married
2 = Domestic partnership
3 = Separated
4 = Divorced
5 = Widowed
6 = Never married
7 = Unknown/Missing
8 = Unable to determine category; see original study variable text

| | | |
|---|---|---|
| field 10 = | Participant's marital status | (Pt_Marital) |
| field 11 = | Mother's marital status | (Mother_Marital) |
| field 12 = | Father's marital status | (Father_Marital) |
| field 13 = | **Original Study Text:** Participant's marital status | (Pt_Marital_orig) |
| field 14 = | **Original Study Text:** Mother's marital status | (Mother_Marital_orig) |
| field 15 = | **Original Study Text:** Father's marital status | (Father_Marital_orig) |
| field 16 = | Participant's years of education (0 to …) | (Pt_Edu_Years) |
| field 17 = | Mother's years of education (0 to …) | (Mother_Edu_Years) |
| field 18 = | Father's years of education (0 to …) | (Father_Edu_Years) |

Codes for **Highest Educational Level CATEGORY Completed** variables:

0 = Never attended or Kindergarten only
1-11 = Grades 1=11
12 = High school, No diploma
13 = GED or equivalent
14 = High school graduate
15 = Completed 12 years of school (but unable to determine if code 12, 13, or 14)
16 = Some college, no degree
17 = Associate degree: occupational/technical/vocational program
18 = Associate degree: academic program (includes non-US post high school/pre college levels)
19 = Associate degree (but unable to determine if code 16 or 17)
20 = Bachelor degree (e.g., BA, AB, BS, BBA)
21 = Completed 16 years of school (but unable to determine if code 16-20)
22 = Master's degree (e.g., MA, MS, MEng, MEd, MBA)
23 = Professional school degree (e.g., MD, DDS, DVM, JD)
24 = Doctoral degree (e.g., PhD, EdD)
25 = Post Master degree (but unable to determine if code 20 or 21)
26 = Unknown/Missing
27 = Unable to determine category from original variable provided

| | | |
|---|---|---|
| field 19 = | Participant's highest educational level category completed | (Pt_Edu_Cat) |
| field 20 = | Mother's highest educational level category completed | (Mother_Edu_Cat) |
| field 21 = | Father's highest educational level category completed | (Father_Edu_Cat) |
| field 22 = | **Original Study Text:** Participant's highest educational level category completed | (Pt_Edu_Cat_orig) |

## Database of Cognitive Training and Remediation– Demographics (DEMO)

field 23 = **Original Study Text:** Mother's highest educational level category completed (Mother_Edu_Cat_orig)
field 24 = **Original Study Text: F**ather's highest educational level category completed (Father_Edu_Cat_orig)

field 25 = **Primary Psychiatric Diagnosis Diagnostic System** (DSM_ICD_code)
1 = Diagnostic and Statistical Manual of Mental Disorders (2000-2012) DSM-IV-TR
2 = Diagnostic and Statistical Manual of Mental Disorders (2013-current) DSM-5
3 = International Classification of Diseases (1990-current) ICD-10
4 = International Classification of Diseases (under development) ICD-11

Codes for **Primary Psychiatric Diagnosis** and **Additional Psychiatric Diagnosis** variables:

1 = Schizophrenia, Unspecified type
2 = Schizophrenia, Disorganized type
3 = Schizophrenia, Paranoid type
4 = Schizophrenia, Residual type
5 = Schizophrenia, Undifferentiated type
6 = Schizophreniform disorder
7 = Schizoaffective disorder
8 = Dysthymia/Persistent Depressive Disorder
9 = Major Depressive Disorder
10 = Bipolar Disorder
11 = Anxiety disorder (GAD, panic disorder, agoraphobia, specific phobia, social phobia, OCD, PTSD)
12 = Cluster A Personality disorder (Paranoid, Schizoid, Schizotypal)
13 = Cluster B or C Personality disorder
14 = Current alcohol or drug abuse or dependence Past alcohol or drug abuse or dependence
15 = Other diagnosis
98 = None (Additional psychiatric diagnoses assessed but not present)
99 = Additional psychiatric diagnoses not reported for study

| | DSM-IV | DSM-5 | ICD-10 |
|---|---|---|---|
| SZ, Unspecified type | 295.XX | 295.9 | F20.9 |
| SZ, Disorganized type | 295.1 | | F20.1 |
| SZ, Paranoid type | 295.3 | | F20.0 |
| SZ, Residual type | 295.6 | | F20.5 |
| SZ, Undifferentiated type | 295.9 | | F20.3 |
| Schizophreniform disorder | 295.4 | 295.4 | F20.81 |
| Schizoaffective disorder | 295.7 | 295.7 | F25.XX |
| Dysthymia / Persistent Depressive Disorder | 300.4 | 300.4 | F34.1 |
| Major Depressive Disorder | 296.XX | 296.XX | F32, F33 |
| Bipolar Disorder | 296.XX | 296.XX | F31 |

field 26 = Primary psychiatric diagnosis (Primary_Dx)
field 27 = Additional psychiatric diagnosis (Other_Dx1)
field 28 = Additional psychiatric diagnosis 2 (Other_Dx2)

**Psychiatric Illness History**

field 29 = Age of onset of psychiatric symptoms (years) (Age_1st_Sx)
field 30 = Age of first treatment for psychiatric symptoms (Age_1st_TxContact)
field 31 = Age at first psychiatric hospitalization (years) (Age_1st_hosp)
field 32 = Total number of psychiatric hospitalizations (Psych_Hosp_Total)
field 33 = Number of psychiatric hospitalizations in last year (Psych_Hosp_Year)

field 34 = Number of sessions completed (Num_Sessions)
field 35 = Percentage of sessions completed (Percent_Sessions)

field 36 = Study completion status (Completer_Status)
**Note**: Investigators' definition of study completion can be found in the Study information database
1 = Completer
2 = Randomized but did not complete any sessions
3 = Non-Completer

# C.4 Medications

## Cognitive Remediation Data Integration – Medication (MED)

field 1 = Study ID number (assigned by NIMH) (Study_ID)
4155 = WYKES1 (n=35)
5693 = BELL (n=77)
6632 = KEEFE (n=53)
7926 = KESHAVAN (n=58)
8134 = WYKES3 (n=40)
9212 = WYKES2 (n=85)
9479 = SILVERSTEIN (n=83)

field 2 = Participant ID number (assigned by NIMH) (Participant_ID)

field 3 = Study condition to which participant was assigned (Study_condition)
1 = Cognitive remediation
2 = Comparison condition
3 = Other

field 4 = Assessment Time Point (Time_point)
5 = Screening
10 = Baseline
21, 22, 23, … = Mid-point 1, 2, 3, …
30 = End-of-treatment
41, 42, 43, … = Follow-up1, 2, 3, …

field 5 = Days since baseline (Days_baseline)

**NOTE: The Days_Baseline variable is calculated using assessment dates when they were provided and calculated from target dates according to the schedule of assessment time points for the study when visits dates were not provided.**

Codes used for psychotropic medications in this database (MED_CODE1 thru MED_CODE5):

1 = aripiprazole
2 = asenapine
3 = chlorpromazine
4 = clozapine
5 = fluphenazine
6 = fluphenazine inj
7 = haloperidol
8 = haloperidol inj
9 = iloperidone
10 = olanzapine
11 = paliperidone
12 = perphenazine
13 = prochlorperazine
14 = promethazine
15 = quetiapine
16 = risperidone
17 = risperidone inj
18 = thiothixene
19 = trifluoperazine
20 = ziprasidone
21 = amitriptyline
22 = bupropion
23 = citalopram
24 = desvenlafaxine
25 = doxepin
26 = duloxetine
27 = escitalopram oxalate
28 = fluoxetine
29 = fluvoxamine
30 = mirtazapine
31 = nortriptyline
32 = paroxetine
33 = sertraline
34 = trazodone
35 = venlafaxine
36 = alprazolam
37 = buspirone
38 = clonazepam
39 = clorazepate
40 = diazepam
41 = eszopiclone
42 = lorazepam
43 = melatonin
44 = oxazepam
45 = phenobarbital
46 = ramelteon
47 = temazepam
48 = zaleplon
49 = zolpidem
50 = zolpidem tartrate
51 = carbamazepine
52 = divalproex
53 = gabapentin
54 = lamotrigine
55 = levetiracetam
56 = lithium
57 = oxcarbazepine
58 = phenytoin
59 = topiramate
60 = valproic acid
61 = zonisamide
62 = amphetamine
63 = amphetamine/dextroamphetamine
64 = atomoxetine
65 = clonidine
66 = dextroamphetamine
67 = modafinil
68 = benztropine mesylate
69 = amantadine
70 = dicyclomine
71 = trihexyphenidyl
72 = Propranolol
73 = Unspecified typical antipsychotic
74 = Unspecified atypical antipsychotic
75 = Dose is for all antipsychotic medications combined in total chlorpromazine equivalents
76 = amisulpiride
77 = sulpiride
78 = droperidol
79 = pipothiazine depot
80 = Flupenthixol depot
81 = zuclopenthixol
82 = zuclopenthixol depot
83 = loxapine
84 = Both (unspecified) typical and (unspecified) atypical meds
85 = None

## Cognitive Remediation Data Integration – Medication (MED)

**Please note that DOSE should be provided regardless of route and FREQ should be provided only for DEPOT (injectable).**

| | | |
|---|---|---|
| field 6 = | Psychotropic Medication (Med1) (from list) | (MED_CODE1) |
| field 7 = | DEPOT or ORAL MED: Med1 Route | (ROUTE1) |
| field 8 = | **ORAL** MED: Med1 dose, MG per day | (DOSE_ORAL_MG1) |
| field 9 = | **DEPOT (injectable)** MED: Med1 dose, CC | (DOSE_INJ_CC1) |
| field 10 = | **DEPOT (injectable)** MED: Med1 frequency | (FREQ_INJ_WK1) |
| field 11 = | Psychotropic Medication (Med2) (from list) | (MED_CODE2) |
| field 12 = | DEPOT or ORAL MED: Med2 Route | (ROUTE2) |
| field 13 = | **ORAL** MED: Med2 dose, MG per day | (DOSE_ORAL_MG2) |
| field 14 = | **DEPOT (injectable)** MED: Med2 dose, CC | (DOSE_INJ_CC2) |
| field 15 = | **DEPOT (injectable)** MED: Med2 frequency | (FREQ_INJ_WK2) |
| field 16 = | Psychotropic Medication (Med3) (from list) | (MED_CODE3) |
| field 17 = | DEPOT or ORAL MED: Med3 Route | (ROUTE3) |
| field 18 = | **ORAL** MED: Med3 dose, MG per day | (DOSE_ORAL_MG3) |
| field 19 = | **DEPOT (injectable)** MED: Med3 dose, CC | (DOSE_INJ_CC3) |
| field 20 = | **DEPOT (injectable)** MED: Med3 frequency | (FREQ_INJ_WK3) |
| field 21 = | Psychotropic Medication (Med4) (from list) | (MED_CODE4) |
| field 22 = | DEPOT or ORAL MED: Med4 Route | (ROUTE4) |
| field 23 = | **ORAL** MED: Med4 dose, MG per day | (DOSE_ORAL_MG4) |
| field 24 = | **DEPOT (injectable)** MED: Med4 dose, CC | (DOSE_INJ_CC4) |
| field 25 = | **DEPOT (injectable)** MED: Med4 frequency | (FREQ_INJ_WK4) |
| field 26 = | Psychotropic Medication (Med5) (from list) | (MED_CODE5) |
| field 27 = | DEPOT or ORAL MED: Med5 Route | (ROUTE5) |
| field 28 = | **ORAL** MED: Med5 dose, MG per day | (DOSE_ORAL_MG5) |
| field 29 = | **DEPOT (injectable)** MED: Med5 dose, CC | (DOSE_INJ_CC5) |
| field 30 = | **DEPOT (injectable)** MED: Med5 frequency | (FREQ_INJ_WK5) |

# C.5 Quality of life, self-esteem and functioning measures

## Database of Cognitive Training and Remediation Studies
## Quality of Life, Self-Esteem and Functioning Measures (QOL FX)

field 1 = Study ID number (assigned by NIMH) (Study_ID)
- 4155 = WYKES1 (n=35)
- 5693 = BELL (n=77)
- 6632 = KEEFE (n=53)
- 7926 = KESHAVAN (n=58)
- 8134 = WYKES3 (n=40)
- 9212 = WYKES2 (n=85)
- 9479 = SILVERSTEIN (n=83)

field 2 = Participant ID number (assigned by NIMH) (Participant_ID)

field 3 = Study condition to which participant was assigned (Study_condition)
- 1 = Cognitive Remediation condition
- 2 = Comparison condition
- 3 = Other

field 4 = Assessment time point (Time_point)
- 5 = Screening
- 10 = Baseline
- 21 = 1st Mid-point, 22 = 2nd mid-point, …,
- 30 = End-of-Treatment
- 41 = 1st follow-up, 42 = 2nd follow-up, …

field 5 = Days since baseline (Days_baseline)

**NOTE: The Days_Baseline variable is calculated using assessment dates when they were provided and calculated from target dates according to the schedule of assessment time points for the study when visits dates were not provided.**

**Rosenberg Self-esteem Scale**: The RSE is a 10-item self-report measure of global self-esteem. The items are scored on a 0-3 scale, but the items that comprise the Confidence factor are worded in a positive direction and the items that comprise the Deprecation factor are worded in a negative direction. The total score is the sum of all of the items, with items in the Deprecation factor reverse scored. Total Scores between 15 and 25 are generally considered to be within normal range; Total Scores below 15 suggest low self-esteem. Data from studies that used rating codes other than 0-3 (e.g., 1-4) have been converted to 0-3 to result in comparable subscale and total scores across studies for these variables.

field 6 = Rosenberg Self-esteem scale: Self-Confirmation factor (items 1, 3, 4, 7, 10; scored 0-3; range 0-15) (RSE_Confirmation)
field 7 = Rosenberg Self-esteem scale: Self-Deprecation factor (items 2, 5, 6, 8, 9; scored 0-3; range 0-15) (RSE_Deprecation)
field 8 = Rosenberg Self-esteem scale: Self-Esteem Total score (items 1-10 scored 0-3; range 0-30) (RSE_Total)

**Rosenberg Self-esteem Scale (alternative scoring)**: For investigators who used an ALTERNATIVE 1-5 scale (1 = strongly agree, 2 = agree, 3 = neither agree nor disagree, 4 = disagree, 5 = strongly disagree), rather than the 0-3 scale in the above variables, the original coding is preserved in these variables.

field 9 = Rosenberg Self-esteem scale: Self-Confirmation factor (items 1, 3, 4, 7, 10; scored 1-5; range 5-25) (RSE_Confirmation_5)
field 10 = Rosenberg Self-esteem scale: Self-Deprecation factor (items 2, 5, 6, 8, 9; scored 1-5; range 5-25) (RSE_Deprecation_5)
field 11 = Rosenberg Self-esteem scale: Self-Esteem Total score (items 1-10; scored 1-5; range 10-50) (RSE_Total_5)

**Heinrichs-Carpenter Quality of Life Scale**: The HCQOL is a 21-item semi-structured interview conducted by a trained clinician to rate quality of life based on patient self-report and professional judgment. Individual items are scored on a 0-6 scale (0 =poor quality of life and 6 =high quality of life) and summed to create factor scores. Data from studies that used other than 0-6 have been converted to 0-6 to result in comparable subscale and total scores across studies. Low Factor Scores suggest a low quality of life.

field 12 = Heinrichs-Carpenter Quality of Life Scale: Interpersonal relations factor (items 1-8 scored 0-6; range 0-48) (HCQOL_Interpersonal_Factor)
field 13 = Heinrichs-Carpenter Quality of Life Scale: Instrumental role factor (items 9-12 scored 0-6; range 0-24) (HCQOL_Role_Factor)
field 14 = Heinrichs-Carpenter Quality of Life Scale: Intrapsychic foundations factor (items 13-17, 20, 21 scored 0-6; range 0-42) (HCQOL_Intrapsychic_Factor)
field 15 = Heinrichs-Carpenter Quality of Life Scale: Common objects and activities factor (items 18-19 scored 0-6; range 0-12) (HCQOL_ObjActivities_Factor)

**Lehman Quality of Life Interview**: The LQLI is a structured interview conducted by a trained non-clinical interviewer to elicit patient ratings of their own quality of life. Although other information is collected with this instrument, the ratings included in DoCTRS are the ratings (or the mean of the ratings) for one (or more) items completed at the end of each section which are scored on a 1-7 scale (1 = terrible, 2 = unhappy, 3 = mostly dissatisfied, 4 = mixed, 5 = mostly satisfied, 6 = pleased, 7 = delighted). The number of items for each section is different in the full version and the brief version, but scores from either version are used for these variables. Data from studies that used anchors other than 1-7 have been converted to 1-7 to result in subscale and total scores that are comparable across studies.

field 16 = Lehman Quality of Life Interview: General life satisfaction (range: 1-7) (LQOL_Life_Satisfaction)
field 17 = Lehman Quality of Life Interview: Living situation (range: 1-7) (LQOL_Living_Sit)
field 18 = Lehman Quality of Life Interview: Daily activities (range: 1-7) (LQOL_Daily_Activ)
field 19 = Lehman Quality of Life Interview: Family relations (range: 1-7) (LQOL_Family_Relations)
field 20 = Lehman Quality of Life Interview: Social relations (range: 1-7) (LQOL_Social_Relations)
field 21 = Lehman Quality of Life Interview: Finances (range: 1-7) (LQOL_Finance)
field 22 = Lehman Quality of Life Interview: Work/school (range: 1-7) (LQOL_Work_School)
field 23 = Lehman Quality of Life Interview: Legal and safety (range: 1-7) (LQOL_LegalSafety)
field 24 = Lehman Quality of Life Interview: Health (range: 1-7) (LQOL_Health)

**Work Behavior Inventory**: The WBI is a 36-item performance assessment instrument. WBI ratings are based on a 15-minute behavioral observation of the worker and a brief interview with the worker's immediate supervisor. Each of the six subscales is comprised of seven items individually assessed on a 5-point continuum (1 = consistently inferior performance and 5 = consistently superior performance) and there is a one-item general work behavior rating (scored using the same 1-5 continuum). Thus, the six individual WBI subscales range from 7 to 35. The summation of the six subscales scores gives the Total score (range: 35 – 245). Data from studies that used anchor coding other than 1-5 have been converted to 1-5 to result in subscale and total scores that are comparable across studies.

| | | |
|---|---|---|
| field 25 = | Work Behavior Inventory: Social skills scale (items A1-A7; range 7-35) | (WBI_Social) |
| field 26 = | Work Behavior Inventory: Cooperativeness scale (items B1-B7; range 7-35) | (WBI_Cooperate) |
| field 27 = | Work Behavior Inventory: Work habits (items C1-C7; range 7-35) | (WBI_Habits) |
| field 28 = | Work Behavior Inventory: Work Quality (items D1-D7; range 7-35) | (WBI_Quality) |
| field 29 = | Work Behavior Inventory: Personal presentation (items E1-E7; range 7-35) | (WBI_Presentation) |
| field 30 = | Work Behavior Inventory: General rating of work behavior (range 1-5) | (WBI_General) |
| field 31 = | Work Behavior Inventory: Total score (items A1-E7; range 35-245) | (WBI_Total) |

**Social Adjustment Scale II**: The SAS-II is a 42- to 54-item scale (depending upon work role questions) that is completed during a 1-hour semi-structured interview. Some items are scored on a 0-4 scale and others are scored on a 0-5 scale, with higher scores indicating worse adjustment. The factor scores included in DoCTRS are derived from the Schooler et al (1976) factor analysis and consist of the means of the items for each factor. Note: Item numbers provided below assume that item 1 (Work Role: Time lost) is numbered 1 and so on, rather than starting numbering at 19 as it is in some versions of the scale. Data from studies that used other than 0-4/0-5 have been converted to 0-4/0-5 to result in comparable subscale and total scores across studies.

| | | |
|---|---|---|
| field 32 = | SAS-II Interpersonal anguish factor (items 4, 5, 22, 23, 24, 25, 27, 28, 29, 38, 44, 45; range 0-4) | (SASII_Anguish) |
| field 33 = | SAS-II Sexual relations factor (items 15, 16, 21, 39, 40, 41; range 0-5) | (SASII_SexRel) |
| field 34 = | SAS-II Primary Relationships factor (items 10, 11, 12, 14; range 0-4 ) | (SASII_Primary) |
| field 35 = | SAS-II Social leisure factor (items 30, 31, 32, 34, 35, 36, 37; range 0-4) | (SASII_SocLei) |
| field 36 = | SAS-II Work affinity factor (items 6, 7; range 0-4) | (SASII_WrkAff) |
| field 37 = | SAS-II Self-care factor (items 26, 42, 43; range 0-4) | (SASII_SelfCare |

**Social Behaviour Scale**: The SBS is a 21-item scale. The anchors/descriptors vary by item but, across items, higher ratings indicate more severe behavioral disruption. The range of ratings varies across items.

| | | |
|---|---|---|
| field 38 = | Social Behaviour Scale: Acting out bizarre ideas (range 0-2) | (SBS_ActOut) |
| field 39 = | Social Behaviour Scale: Ability to make appropriate social contacts (range 0-4) | (SBS_SocialCon) |
| field 40 = | Social Behaviour Scale: Attention-seeking behavior (range 0-4) | (SBS_AttnSeek) |
| field 41 = | Social Behaviour Scale: Coherence of conversation (range 0-4) | (SBS_Coherence) |
| field 42 = | Social Behaviour Scale: Concentration (range 0-2) | (SBS_Concentration) |
| field 43 = | Social Behaviour Scale: Depression (range 0-3) | (SBS_Depression) |
| field 44 = | Social Behaviour Scale: Hostility/friendliness (range 0-3) | (SBS_HostilityFriendliness) |
| field 45 = | Social Behaviour Scale: Personal appearance and hygiene (range 0-4) | (SBS_AppearanceHygiene) |
| field 46 = | Social Behaviour Scale: Taking initiative (range 0-4) | (SBS_Initiative) |

| | | |
|---|---|---|
| field 47 = | Social Behaviour Scale: Laughing and talking to self (range 0-4) | (SBS_Laughing) |
| field 48 = | Social Behaviour Scale: Socially unacceptable manners or habits (range 0-4) | (SBS_MannersHabits) |
| field 49 = | Social Behaviour Scale: Oddity/Inappropriateness of conversation (range 0-4) | (SBS_Inapp_Conversation) |
| field 50 = | Social Behaviour Scale: Other behaviors that impede progress (range 0-4) | (SBS_Other_Behaviors) |
| field 51 = | Social Behaviour Scale: Overactivity and restlessness (range 0-4) | (SBS_Overactivity) |
| field 52 = | Social Behaviour Scale: Panic attacks and phobias (range 0-4) | (SBS_PanicPhobias) |
| field 53 = | Social Behaviour Scale: Posturing and mannerisms (range 0-4) | (SBS_Posturing) |
| field 54 = | Social Behaviour Scale: Inappropriate sexual behavior (range 0-4) | (SBS_Inapp_Sexual) |
| field 55 = | Social Behaviour Scale: Slowness (range 0-4) | (SBS_Slowness) |
| field 56 = | Social Behaviour Scale: Suicidal ideas and self-harm (range 0-4) | (SBS_Suicidal) |
| field 57 = | Social Behaviour Scale: Underactivity (range 0-4) | (SBS_Underactivity) |
| field 58 = | Social Behaviour Scale: Violent, threatening or destructive behavior (range 0-3) | (SBS_Violent) |
| field 59 = | Social Behaviour Scale: total score Q1-21 | (SBS_Total) |

### *Scoring Conventions for Factors and Total Scores*

In order to standardize scoring conventions across studies, Factors and Total score for all forms were computed according to the following rules:

***Factor* Score MEANS:**

Factor MEANS were computed by summing the appropriate items for that Factor and then dividing by the number of items. To accommodate missing data, the denominator was adjusted to include only those items that were non-missing. That is, the Factor Score is computed as the mean of those items present. If less than 80% of the items were present for a factor, that Factor Score Mean was not computed.

***Factor* Score TOTALS:**

Factor TOTALS were computed by multiplying the Factor Mean (computed according to the rules described above) by the number of items appropriate for that Factor. Thus, the Factor Score Total is "pro-rated" to account for missing data. Note, if less than 80% of the items were present for a factor, that Factor Score Total was not computed.

***Total* Score MEANS:**

Total Score MEANS were computed by summing all items in the scale and then dividing by the number of items. To accommodate missing data, the denominator was adjusted to include only those items that were non-missing. That is, the total score is computed as the mean of those items present. If less than 80% of the items were present **or** if any factor was missing, the Total Score Mean was not computed.

***Total* Score TOTALS:**

Total Score TOTALS were computed by multiplying the Total Score Mean (computed according to the rules described above) by the number of items in the scale. Thus, the Total Score Total is "pro-rated" to account for missing data. Note, if less than 80% of the items were present **or** if any factor was missing, the Total Score Total was not computed.

If data from individual items were not provided, any Factor and/or Total scores were not re-computed.

# C.6 Symptom data

## Database of Cognitive Training and Remediation Studies – Symptom data (SXS)

field 1 = Study ID number (assigned by NIMH) (Study_ID)
- 4155 = WYKES1 (n=35)
- 5693 = BELL (n=77)
- 6632 = KEEFE (n=53)
- 7926 = KESHAVAN (n=58)
- 8134 = WYKES3 (n=40)
- 9212 = WYKES2 (n=85)
- 9479 = SILVERSTEIN (n=83)

field 2 = Participant ID number (assigned by NIMH) (Participant_ID)

field 3 = Study condition to which participant was assigned (Study_condition)
- 1 = Cognitive Remediation condition
- 2 = Comparison condition
- 3 = Other

field 4 = Assessment time point (Time_point)
- 5 = Screening
- 10 = Baseline
- 21, 22, 23, … = Mid-point 1, 2, 3, …
- 30 = End-of-treatment
- 41, 42, 43, … = Follow-up1, 2, 3, …

field 5 = Days since baseline (Days_baseline)

**NOTE: The Days_Baseline variable is calculated using assessment dates when they were provided and calculated from target dates according to the schedule of assessment time points for the study when visits dates were not provided.**

**Notes regarding calculation of total and factor scores and missing data conventions:**

In order to insure accuracy and consistency, total scores and factor scores on the PANSS and BPRS were (re-)computed by DoCTRS. However, if item-level data were not provided, the investigator-provided total scores and factors scores were retained.

Conventions used for handling missing data in computing factor and total scores for the PANSS and BPRS: Within a factor, data from 80% of the items must be present in order to calculate the score for that factor. If a factor has data missing for some items but at least 80% of the items are present, the mean of the available scores within the factor was used to replace the missing item(s) value(s). The replaced value(s) are used to calculate factor scores and total scores.

Total scores were calculated only for subjects for whom (1) ratings were available on at least 80% of the items in the full scale and (2) all factors were able to be scored. Therefore, if any individual factor could not be scored due to less than 80% of the items in that factor being present, then the overall summary score would not be calculated. If the overall summary score could be calculated based on above rules, any missing scores were assigned the value that they were assigned within their own factor.

**Positive and Negative Syndrome Scale (PANSS)**

The PANSS (Kay SR, Fiszbein A, Opler LA: The Positive and Negative Syndrome Scale (PANSS) for schizophrenia. Schizophrenia Bulletin 13:261-276, 1987a) is a 30-item rating scale that is specifically developed to assess individuals with schizophrenia.

Codes for the 30 PANSS Items (P1-P7, N1-N7, G1-G16):

1 = Absent
2 = Minimal
3 = Mild
4 = Moderate
5 = Moderate-severe
6 = Severe
7 = Extreme

| | | |
|---|---|---|
| field 6 = | PANSS item P1 Delusions | (PANSS_delu) |
| field 7 = | PANSS item P2 Conceptual Disorganization | (PANSS_conc) |
| field 8 = | PANSS item P3 Hallucinatory behavior | (PANSS_hall) |
| field 9 = | PANSS item P4 Excitement | (PANSS_exci) |
| field 10 = | PANSS item P5 Grandiosity | (PANSS_gran) |
| field 11 = | PANSS item P6 Suspiciousness | (PANSS_susp) |
| field 12 = | PANSS item P7 Hostility | (PANSS_host) |
| field 13 = | PANSS item N1 Blunted Affect | (PANSS_blun) |
| field 14 = | PANSS item N2 Emotional Withdrawal | (PANSS_emot) |
| field 15 = | PANSS item N3 Poor Rapport | (PANSS_rapp) |
| field 16 = | PANSS item N4 Passive/Apathetic Social Withdrawal | (PANSS_apath) |
| field 17 = | PANSS item N5 Difficulty in Abstract Thinking | (PANSS_abst) |
| field 18 = | PANSS item N6 Lack of Spontaneity | (PANSS_spont) |
| field 19 = | PANSS item N7 Stereotyped Thinking | (PANSS_ster) |
| field 20 = | PANSS item G1 Somatic Concern | (PANSS_somc) |
| field 21 = | PANSS item G2 Anxiety | (PANSS_anxi) |
| field 22 = | PANSS item G3 Guilt Feelings | (PANSS_guil) |
| field 23 = | PANSS item G4 Tension | (PANSS_tens) |
| field 24 = | PANSS item G5 Mannerisms and Posturing | (PANSS_mann) |
| field 25 = | PANSS item G6 Depression | (PANSS_depr) |
| field 26 = | PANSS item G7 Motor Retardation | (PANSS_motr) |
| field 27 = | PANSS item G8 Uncooperativeness | (PANSS_unco) |
| field 28 = | PANSS item G9 Unusual Thought Content | (PANSS_unus) |
| field 29 = | PANSS item G10 Disorientation | (PANSS_diso) |
| field 30 = | PANSS item G11 Poor Attention | (PANSS_attn) |
| field 31 = | PANSS item G12 Lack of Judgment and Insight | (PANSS_judg) |
| field 32 = | PANSS item G13 Disturbance of Volition | (PANSS_voli) |
| field 33 = | PANSS item G14 Poor Impulse control | (PANSS_impu) |
| field 34 = | PANSS item G15 Preoccupation | (PANSS_preo) |
| field 35 = | PANSS item G16 Active Social Avoidance | (PANSS_soca) |

PANSS summary scores:

| | | |
|---|---|---|
| field 36 = | PANSS Total (Sum of all 30 items; Range: 30-210) | (PANSS_total) |

field 37 = PANSS Positive Score (Sum of items P1-P7; Range: 7-49) (PANSS_posit)
field 38 = PANSS Negative Score (Sum of items N1-N7; Range: 7-49) (PANSS_nega)
field 39 = PANSS General Score (Sum of items G1-G16; Range: 16-112)(PANSS_genr)

PANSS factor scores:
The reference for these factor scores is: Lindenmayer J-P, Bernstein-Hyman R, and Grochowski S: A new five factor model of schizophrenia. Psychiatric Quarterly 65:299-322, 1994.

There are alternative factor solutions that users may calculate from the item-level data in the database (See Wallwork et al.: Searching for a consensus five-factor model of the Positive and Negative Syndrome Scale for schizophrenia. Schizophrenia Research 137: 246-250, 2012).

field 40 = PANSS Negative Factor (N1+N2+N3+N4+N6+G16; range 6-42) (PANNS_neg_factor)
field 41 = PANSS Excitement Factor (P4+P7+G4+G14; range 4-28) (PANNS_excite_factor)
field 42 = PANSS Cognitive Factor (P2+N5+G5+G10+G11; range 5-35) (PANNS_cog_factor)
field 43 = PANSS Positive Factor (P1+P5+P6+G9; range 4-28) (PANNS_pos_factor)
field 44 = PANSS Depression component (G1+G2+G3+G6+G15, range 5-35) (PANNS_dep_factor)

**Brief Psychiatric Rating Scale (BPRS)**
The BPRS (Overall, JE and Gorham, DR, The Brief Psychiatric Rating Scale, Psychol Rep, 10:799-812, 1962) measures psychotic symptoms and was first published in 1962. The scale provides evaluation of treatment response in both clinical drug trials and routine clinical settings. The standard version of the BPRS contains 18 items and the expanded versions contain up to 24 items.

Note: The original 18 BPRS items are indicated by asterisk (*).

Codes for the 24 BPRS Items:
1 = Not present
2 = Very mild
3 = Mild
4 = Moderate
5 = Moderately severe
6 = Severe
7 = Extremely severe

field 45 = BPRS item 1 Somatic concern * (BPRS_somc)
field 46 = BPRS item 2 Anxiety * (BPRS_anxi)
field 47 = BPRS item 3 Depressive Mood * (BPRS_depr)
field 48 = BPRS item 4 Guilt Feelings* (BPRS_guil)
field 49 = BPRS item 5 Hostility * (BPRS_host)
field 50 = BPRS item 6 Suspiciousness * (BPRS_susp)
field 51 = BPRS item 7 Unusual thought content * (BPRS_unus)
field 52 = BPRS item 8 Grandiosity * (BPRS_gran)
field 53 = BPRS item 9 Hallucinatory Behavior * (BPRS_hall)

| | | | |
|---|---|---|---|
| field 54 = | BPRS item 10 | Disorientation * | (BPRS_diso) |
| field 55 = | BPRS item 11 | Conceptual disorganization * | (BPRS_conc) |
| field 56 = | BPRS item 12 | Excitement * | (BPRS_exci) |
| field 57 = | BPRS item 13 | Motor retardation * | (BPRS_motr) |
| field 58 = | BPRS item 14 | Blunted affect * | (BPRS_blun) |
| field 59 = | BPRS item 15 | Tension * | (BPRS_tens) |
| field 60 = | BPRS item 16 | Mannerisms and posturing * | (BPRS_mann) |
| field 61 = | BPRS item 17 | Uncooperativeness * | (BPRS_unco) |
| field 62 = | BPRS item 18 | Emotional withdrawal * | (BPRS_emot) |
| field 63 = | BPRS item 19 | Suicidality | (BPRS_suic) |
| field 64 = | BPRS item 20 | Self neglect | (BPRS_self) |
| field 65 = | BPRS item 21 | Bizarre behavior | (BPRS_bizb) |
| field 66 = | BPRS item 22 | Elated mood concern | (BPRS_elat) |
| field 67 = | BPRS item 23 | Motor hyperactivity | (BPRS_mohy) |
| field 68 = | BPRS item 24 | Distractibility | (BPRS_distr) |

BPRS Summary Scores

BPRS 18-Item Total score (BPRS_total)

The reference for the following factor solution is: Overall et al. Major psychiatric disorders: A four-dimensional model. Archives of General Psychiatry 16: 146-151, 1967.

This factor solution has good empirical support, however there are alternative factor solutions that users may calculate from the data items provided in the database; See Shafer A. Meta-analysis of the Brief Psychiatric Rating Scale factor structure, Psychological Assessment 17: 324-335, 2005.

| | | |
|---|---|---|
| field 69 = | BPRS Factor score: Anxiety and depression (Sum of items 2, 3 & 4; range 3-21) | (BPRS_anx_factor) |
| field 70 = | BPRS Factor score: Hostility and suspiciousness (Sum of items 5, 6 & 17; range 3-21) | (BPRS_host_factor) |
| field 71 = | BPRS Factor score: Thought disturbance (Sum of items 7, 9 & 11; range 3-21) | (BPRS_thought_factor) |
| field 72 = | BPRS Factor score: Withdrawal and retardation (Sum of items 13, 14 & 18; range 3-21) | (BPRS_withd_factor) |

**Scale for the Assessment of Negative Symptoms (SANS)**
References:
Andreasen, NC, Scale for the Assessment of Negative Symptoms: SANS, Iowa, the University of Iowa, 1981.
Andreasen, NC, Negative symptoms in schizophrenia: definition and reliability, Arch Gen Psychiatry 39:784-788, 1982.

This 25-item scale measures the following 5 domains: Affective flattening, alogia, avolition/apathy, anhedonia/asociality, and attention. There are individual items within each domain followed by the rater's global assessment of the domain (in bold below).

**Scale for the Assessment of Positive Symptoms (SAPS)**
Reference: Andreasen, NC, Scale for the assessment of positive symptoms: SAPS, Iowa City, IA, University of Iowa, 1984.

## Database of Cognitive Training and Remediation Studies – Symptom data (SXS)

This 34-item scale measures the following 4 domains: Hallucinations, Delusions, Bizarre behavior, and Positive/formal thought disorder. There are individual items for each domain followed by the rater's global assessment of the domain.

Note: The SANS and the SAPS are listed together in the documentation because when used in studies of schizophrenia, they are often used together.

Codes for the SANS and SAPS Items:

0 = Not at all
1 = Questionable
2 = Mild
3 = Moderate
4 = Marked
5 = Severe

| | | |
|---|---|---|
| field 73 = | SANS item 1 Unchanging facial expression | (SANS1_express) |
| field 74 = | SANS item 2 Decreased spontaneous movements | (SANS2_movement) |
| field 75 = | SANS item 3 Paucity of expressive gestures | (SANS3_gestures) |
| field 76 = | SANS item 4 Poor eye contact | (SANS4_eye) |
| field 77 = | SANS item 5 Affective non-responsivity | (SANS5_respons) |
| field 78 = | SANS item 6 Lack of vocal inflections | (SANS6_vocal) |
| **field 79 =** | **SANS item 7 Global rating of affective flattening** | **(SANS7_global_flattening)** |
| field 80 = | SANS item 8 Inappropriate affect | (SANS8_inapprop) |
| field 81 = | SANS item 9 Poverty of speech | (SANS9_povspeech) |
| field 82 = | SANS item 10 Poverty of content | (SANS10_povcontent) |
| field 83 = | SANS item 11 Blocking | (SANS11_blocking) |
| field 84 = | SANS item 12 Increased latency of response | (SANS12_latency) |
| **field 85 =** | **SANS item 13 Global rating of alogia** | **(SANS13_global_alogia)** |
| field 86 = | SANS item 14 Grooming and Hygiene | (SANS14_hygiene) |
| field 87 = | SANS item 15 Impersistence at work or school | (SANS15_impersist) |
| field 88 = | SANS item 16 Physical anergia | (SANS16_anergia) |
| **field 89 =** | **SANS item 17 Global rating of avolition-apathy** | **(SANS17_global_avolition)** |
| field 90 = | SANS item 18 Recreational interests and activities | (SANS18_recreate) |
| field 91 = | SANS item 19 Sexual interest and activity | (SANS19_sexual) |
| field 92 = | SANS item 20 Ability to feel intimacy and closeness | (SANS20_intimacy) |
| field 93 = | SANS item 21 Relationships with friends and peers | (SANS21_friends) |
| **field 94 =** | **SANS item 22 Global rating of anhedonia-asociality** | **(SANS22_global_asociality)** |
| field 95 = | SANS item 23 Social inattentiveness | (SANS23_socinattention) |
| field 96 = | SANS item 24 Inattentiveness during mental status testing | (SANS24_statinattention) |
| **field 97 =** | **SANS item 25 Global rating of attention** | **(SANS25_global_attention)** |

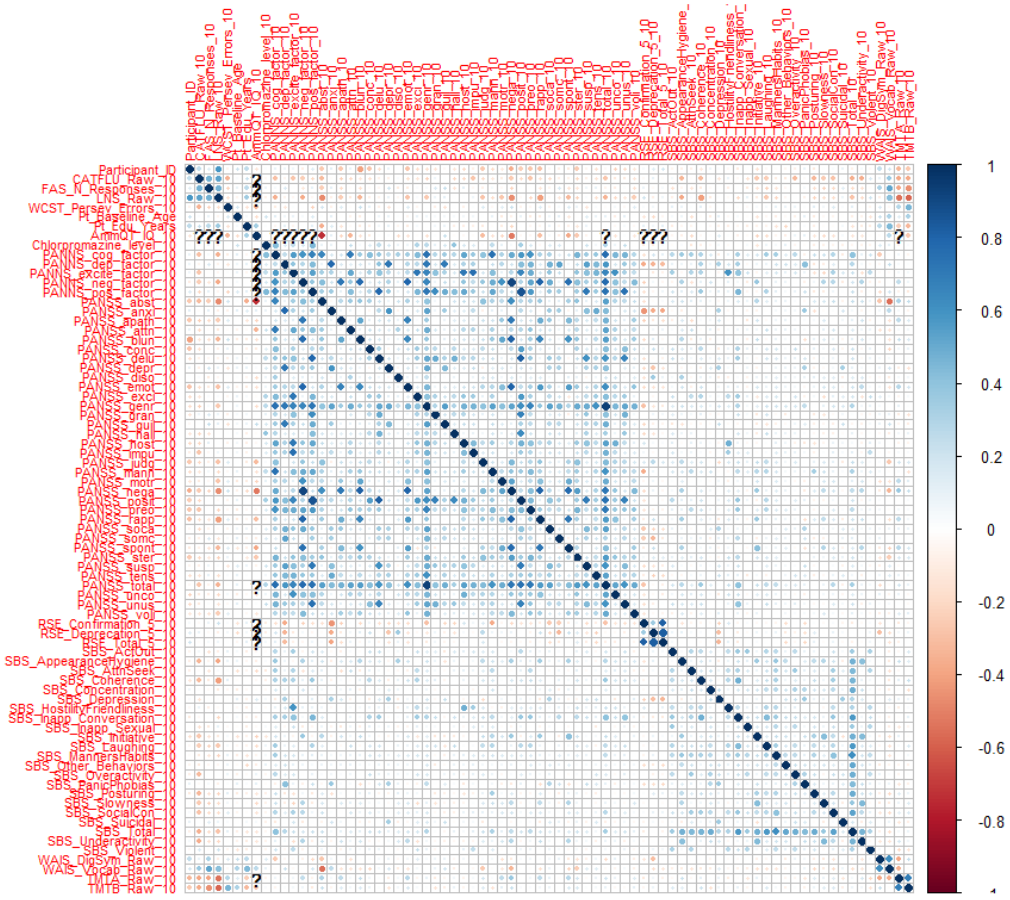
field 98 = SAPS item 1 Auditory hallucinations (SAPS1_auditory)
field 99 = SAPS item 2 Voices commenting (SAPS2_comment)
field 100 = SAPS item 3 Voices conversing (SAPS3_converse)
field 101 = SAPS item 4 Somatic or tactile hallucinations (SAPS4_somatic)
field 102 = SAPS item 5 Olfactory hallucinations (SAPS5_olfactory)
field 103 = SAPS item 6 Visual hallucinations (SAPS6_visual)
**field 104 = SAPS item 7 Global rating of severity of hallucinations (SAPS7_global_hall)**
field 105 = SAPS item 8 Persecutory delusions (SAPS8_persec)
field 106 = SAPS item 9 Delusions of jealousy (SAPS9_jealous)
field 107 = SAPS item 10 Delusions of sin or guilt (SAPS10_guilt)
field 108 = SAPS item 11 Grandiose delusions (SAPS11_grand)
field 109 = SAPS item 12 Religious delusions (SAPS12_religious)
field 110 = SAPS item 13 Somatic delusion (SAPS13_somaticdel)
field 111 = SAPS item 14 Delusions of references (SAPS14_reference)
field 112 = SAPS item 15 Delusions of being controlled (SAPS15_control)
field 113 = SAPS item 16 Delusions of mind reading (SAPS16_mind)
field 114 = SAPS item 17 Delusions of thought broadcasting (SAPS17_broadcast)
field 115 = SAPS item 18 Delusions of thought insertion (SAPS18_insertion)
field 116 = SAPS item 19 Delusions of thought withdrawal (SAPS19_withdrawal)
**field 117 = SAPS item 20 Global rating of severity of delusions (SAPS20_global_del)**
field 118 = SAPS item 21 Clothing and appearance (SAPS21_appearance)
field 119 = SAPS item 22 Social and sexual behavior (SAPS22_social)
field 120 = SAPS item 23 Aggressive and agitated behavior (SAPS23_aggress)
field 121 = SAPS item 24 Repetitive or stereotyped behavior (SAPS24_repetitive)
**field 122 = SAPS item 25 Global rating of severity of bizarre behavior (SAPS25_global_biz)**
field 123 = SAPS item 26 Derailment (SAPS26_derailment)
field 124 = SAPS item 27 Tangentiality (SAPS27_tangent)
field 125 = SAPS item 28 Incoherence (SAPS28_incoherence)
field 126 = SAPS item 29 Illogicality (SAPS29_illogicality)
field 127 = SAPS item 30 Circumstantiality (SAPS30_circumstan)
field 128 = SAPS item 31 Pressured speech (SAPS31_pressured)
field 129 = SAPS item 32 Distractible speech (SAPS32_distract)
field 130 = SAPS item 33 Clanging (SAPS33_clanging)
**field 131 = SAPS item 34 Global rating of positive formal thought disorder (SAPS34_global_thought)**

# Appendix D

# Prediction models results

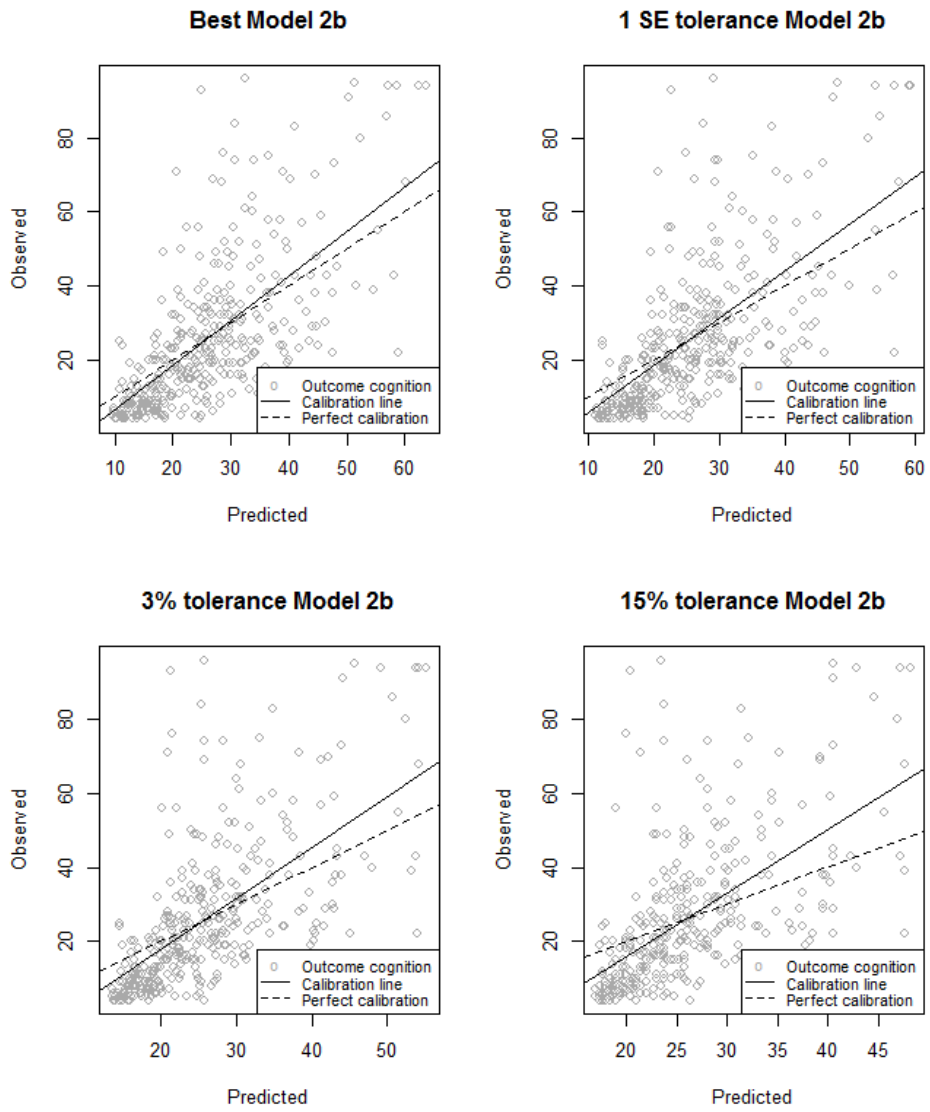
## D.1 Plot of the correlation matrix of the potential predictors used to develop the prediction models

Figure D.1: Plot of the correlation matrix of the potential predictors: Correlations were computed using pairwise complete observations. Because of the percentage of missing data they might underestimate the true correlations. Positive correlations are displayed in blue and negative correlations in red colour. Colour intensity and the size of the circle are proportional to the correlation coefficients.



## D.2 Results for the precision medicine Models 2a and 2b with WCST PE as outcome

Figure D.2: **Model 2b** predictions versus observed outcome values for the best $\lambda$ model, the one SE, the 3% and the 15% tolerance models. Apparent calibration lines are shown.



| Covariate | Model 3 | |
|---|---|---|
| | Uncalibrated coef. | Re-calibrated coef. |
| Intercept | 1.5069 | 1.7386 |
| Education category, primary or less vs 'other' | -0.0584 | -0.0689 |
| log(TMTA) | -0.2640 | -0.3063 |
| log(TMTB) | -0.1820 | -0.2116 |
| Cognition | 0.4973 | 0.5728 |
| Errorless learning technique rank order, No central vs 'other' | 0.0257 | 0.0283 |
| Session duration, 90 vs 60 minutes | 0.0056 | 0.0050 |
| Verbal memory target rank order, No priority target vs 'other' | 0.0001* | -0.0014 |
| CR sessions delivered one-on-one, yes vs no | -0.0001* | -0.0014 |
| Target follow-up, 24 weeks vs 'other' | -0.0398 | -0.0474 |
| CATFLU | 0.0014 | 0.0002 |
| LNS | 0.0357 | 0.0398 |

Table D.2: Final (3% tolerance) **Model 3** uncalibrated and re-calibrated coefficients (coef.). The word 'other' was used to indicate the union of a categorical variable levels for which the dummy was not selected. The colon ':' indicates an interaction. The star sign * means that the estimates were less than $|10^{-4}|$ in absolute value: $0<0.0001^*<0.0001$ and $-0.0001<-0.0001^*<0$.

| Covariate | Model 2a | | Model 2b | |
|---|---|---|---|---|
| | Uncalibrated coef. | Re-calibrated coef. | Uncalibrated coef. | Re-calibrated coef. |
| Intercept | 15.2644 | 10.9927 | 17.0385 | 13.7781 |
| AQT for IQ | -0.0272 | -10.5195 | -0.0415 | -9.8796 |
| PANSS abstract thinking (N5) | 0.0686 | -10.3847 | 0.1454 | -9.6207 |
| Target duration of CRT (weeks) | | | -0.0098 | -9.8358 |
| Target follow-up, 24 weeks vs 'other' | 1.3634 | -8.5631 | 2.3948 | -6.5052 |
| WCST PE | 0.4282 | -9.8788 | 0.4018 | -9.2657 |

Table D.1: Final (3% tolerance) **Model 2a and 2b** uncalibrated and re-calibrated coefficients (coef.). The word 'other' was used to indicate the union of a categorical variable levels for which the dummy was not selected. The colon ':' indicates an interaction.

## D.3 Prognostic Models 3, 4a and 4b: results

| Covariate | Model 4a | | Model 4b | |
|---|---|---|---|---|
| | Uncalibrated coef. | Re-calibrated coef. | Uncalibrated coef. | Re-calibrated coef. |
| Intercept | 11.7729 | 6.3394 | 15.9155 | 12.3397 |
| Education category, primary or less vs 'other' | 1.4586 | 1.4586 | 1.8128 | -7.0249 |
| AQT for IQ | | | -0.0357 | -9.5632 |
| PANSS abstract thinking (N5) | 0.2038 | -9.7328 | 0.0418 | -9.4568 |
| PANSS Negative Score | | | 0.0519 | -9.4429 |
| Target duration of CRT (weeks) | | | -0.0153 | -9.5352 |
| Target follow-up, 24 weeks vs 'other' | 1.6994 | -7.9896 | 2.3903 | -6.2320 |
| WCST PE | 0.4358 | -9.4105 | 0.3982 | -8.9674 |

Table D.3: Final (3% tolerance) **Model 4a and 4b** uncalibrated and re-calibrated coefficients (coef.). The word 'other' was used to indicate the union of a categorical variable levels for which the dummy was not selected. The colon ':' indicates an interaction.